Le contournement de résistance par Melampsora Larici-populina l'agent de la rouille du peuplier : impact démographique et déterminisme génétique

Antoine Persoons

Faculté des Sciences et Technologies
École doctorale Ressources Procédés Produits Environnement
Département de Formation Doctorale Sciences Agronomiques et Forestières, Biologie et Écologie, Biotechnologies
Spécialité Biologie Végétale et Forestière

Thèse

Présentée pour l'obtention du titre de
Docteur de l'Université de Lorraine
par

Antoine Persoons

Les contournements de résistance par *Melampsora larici-populina,* l'agent de la rouille du peuplier : impact démographique et déterminisme génétique

Soutenue publiquement le 15 décembre 2015 devant la commission d'examen :

Martin LASCOUX	Professeur, Université d'Uppsala	Rapporteur
Christophe LEMAIRE	Maitre de Conférence, Université d'Angers	Rapporteur
Éric GELHAYE	Professeur, Université de Lorraine	Examinateur
Pierre GLADIEUX	Chargé de recherche, INRA Montpellier	Examinateur
Sébastien DUPLESSIS	Directeur de recherche, INRA Nancy	Directeur de thèse
Stéphane DE MITA	Chargé de recherche, INRA Nancy	Co-Directeur

Membre invité :

Fabien HALKETT	Chargé de recherche, INRA Nancy	Co-Directeur

UMR 1136 INRA / Université de Lorraine, Interactions Arbres - Microorganismes
Centre INRA de Nancy, Champenoux

Remerciements

La meilleure métaphore que j'ai trouvée pour décrire mon expérience de la thèse est celle d'une grossesse. Apprenant la bonne nouvelle, je me suis lancé plein d'entrain dans cette aventure il y a maintenant 3 ans et 1 mois (soit la durée de gestation du requin lézard, un animal solitaire des grandes profondeurs, de là à y voir une coïncidence...). Rapidement, les premiers signes de mon état se font ressentir, les grandes périodes de joie succèdent aux moments de désarroi mais, sentant le petit scientifique grandir en moi, je m'accroche et persévère. Soudain en un souffle, sans même y penser, me voilà proche du terme, proche et en même temps si loin. L'accouchement est long et pénible (« c'est toujours le cas pour un premier » ; Poncif and Platitude, 2015). Mais l'alternance de stress, de doute et d'envie de tout arrêter prend fin ici et maintenant ; ça y est, elle est là ma petite thèse. Certes, elle est encore un peu fripée et sale mais un peu de mise en page et il n'y paraitra plus. Ainsi, avant de présenter mon bébé à la famille et à la communauté, il me faut remercier ceux qui ont rendu ce petit miracle biologique possible.

Tout d'abord les papas (« 3 papas ? Mais dans quel monde vit on ? » ; Morano, 2015). Fabien, tu m'as appris le professionnalisme inhérent à tout travail scientifique et à peu près tout ce que je sais de la génétique des populations (et je partais de loin !). Merci pour ton didactisme et ta bienveillance. Stéphane, tu as réussi l'exploit de m'initier à la bioinformatique et à la coalescence. Merci pour ta patience (il en a fallu), ton humour (ça c'est vraiment ...) et ton implication sans faille dans mes travaux. Seb, tu m'as montré ce qu'être chercheur implique, les bons (publications, congrès, analyses) et les moins bons côtés (réunions, administrations, recherche de financement) en me prodiguant toujours tes conseils avisés. Tu as toujours été à mes côtés et je garde un souvenir mémorable de nos nombreux fous rires pendant les pauses. Je pense sérieusement avoir bénéficié d'un encadrement hors normes pour mes travaux de thèse (3 spécialistes, 3 disciplines différentes, deux équipes). Vous m'avez fait grandir en même temps que mes travaux, pour tout ce que ça implique et au-delà un grand merci à vous trois.

Ensuite, la « famille scientifique » (la grande et belle famille... enfin « grande » en tout cas). La diversité qui règne au sein de l'équipe écologie est à faire pâlir d'envie les bactéries du sol. Axelle et Olivier (impossible de vous nommer séparément), vous êtes la pierre angulaire de cette équipe. Votre complicité et votre bonne humeur constante sont un vrai plaisir jour après jour au labo. Un merci particulier à toi Axelle pour l'aide précieuse que tu m'as fournie sur la biologie moléculaire. Jérémy, ta passion pour l'ASNL n'a d'égale que ton dédain des Messins, j'adore ton caractère entier, ne change rien et merci pour toutes les discussions footballistiques autour d'un café. Anaïs merci pour ta bonne humeur et les discussions de couloir, je te souhaite bon courage avec ta mycothèque ! Bénédicte je garde un souvenir impérissable des explosions de paillettes sorties de l'azote liquide, je n'ai jamais autant sursauté de ma vie. Merci à Claude, Benoit, Pascal et François

pour les grandes discussions scientifiques aux pauses, vous m'avez forgé une culture scientifique et un sens critique indispensable. Merci à Marius Colin que j'ai encadré lors de son stage de Master 1, tu as continué en science je n'ai donc pas dû être trop mauvais, je te souhaite bon courage pour la suite. Je remercie également Diane Saunders de m'avoir accepté en post-doc, ça a un peu précipité cette thèse mais j'ai vraiment hâte de rejoindre votre équipe et de travailler sur ce sujet passionnant dès janvier.

Les frères et sœurs de doctorats avec qui on a partagé cette aventure. Michaël tout d'abord, on aura quasiment fait nos thèses ensemble puisque tu as soutenu 6 mois avant moi. Merci mon pote de bureau, tu as toujours été là pour moi et inversement je l'espère. J'ai trouvé en toi un véritable ami avec qui j'ai partagé les plus grands délires, des parties de ping-pong endiablées, plus de 3000 pauses clopes (si si...), des congrès arrosés et j'en passe. Je te souhaite vraiment le meilleur pour ton poste au CIRAD, tu le mérites. Jaime, on s'est croisé pendant un an, je commençais et tu terminais, un grand merci de m'avoir pris sous ton aile. Aurore merci pour ta gentillesse, tes rires et ton accent campussien (ça m'a permis d'apprendre ce mot). Marie fraichement arrivée, je te souhaite bon courage pour ta thèse je suis sûr que tu vas cartonner. Enfin, merci aux doctorants d'éco-génomique que j'ai pu croiser et avec qui je me suis très bien entendu : merci à Alice, Vincent, Thibault, Stéphane, Cora, Henri, Clément et Ben que j'ai hâte d'aller retrouver à Norwich.

Ensuite, la famille non scientifique (ceux qui vivent une vraie vie quoi). Un grand merci à ma mère, tes petits plats du week-end associés à ton intérêt constant pour mon travail m'ont toujours remonté le moral, je peux désormais me vanter d'avoir la mère journaliste la plus calée en génétique des populations ! Un énorme merci à toi pour tout ce que tu représentes. Mes frères et sœurs, merci à Alice, Ariane et Augustin pour les supers réunions de famille qui m'ont toujours changé les idées, merci pour votre soutien. Ensuite, Sylvie et Didier et globalement toute la « smala Duguet », vous m'avez intégré à votre famille sans aucune retenue et m'avez toujours soutenu dans les bons comme dans les mauvais moments, et ce depuis mes 17 ans. Un grand merci à vous, je ne pourrai jamais vous rendre ce que vous m'avez offert. Merci à mes potes PL, Eric, Typhène, Loic, Igor pour toutes les soirées, balades et voyages qui m'ont sorti de mon quotidien. Et merci à notre grand couple d'amis nancéens, Lili et Olivier pour toutes les soirées que nous avons passées ensemble, mille mercis pour votre amitié sans failles.

Pour terminer, bien sûr Marie, devenue mon épouse au cours de cette thèse. Merci d'avoir été présente du premier au dernier moment, d'avoir supporté mon caractère en fin de rédaction comme lors des matchs de la Belgique. Tu es mon indispensable pilier et la source de ma volonté sans qui rien de tout ça n'aurait été possible. Je t'aime.

Merci à tous

SOMMAIRE

du peuplier *Melampsora larici-populina*

Chapitre 5 - Synthèse et conclusion générale

Chapitre 1

Introduction générale

Chapitre 1 - Introduction générale et synthèse bibliographique

Homo sapiens a acquis son statut d'espèce sédentaire il y a 10 000 ans, avec la naissance de l'agriculture dans le croissant fertile du Moyen-Orient. Cette révolution a engendré des impacts énormes sur notre espèce tels que la naissance des sociétés lors de la sédentarisation, la grande capacité de conservation des aliments et l'allongement de la vie dû à une alimentation plus riche en glucides (à long terme, car à court terme les maladies et l'alimentation moins riche en vitamines ont engendré une baisse de l'espérance de vie). En contrepartie, elle a apporté quelques méfaits tels que le développement de maladies infectieuses dues à la hausse de la densité de population hôtes, ainsi que l'augmentation significative du nombre d'infections dentaires. L'expansion de l'humanité est historiquement liée à ses capacités agricoles, ce qui est toujours d'actualité dans le contexte actuel d'explosion démographique (10 milliards d'êtres humains en 2050) (figure1). Les agents pathogènes posent des problèmes aux agriculteurs depuis que les premières plantes cultivées ont été domestiquées (Balter, 2007). Beaucoup des espèces cultivées les plus importantes d'aujourd'hui sont originaires du Croissant fertile au Moyen-Orient. Les huit plus importantes cultures fondatrices du Néolithique dans le Croissant fertile étaient les progéniteurs sauvages du blé, engrain, orge, lin, pois chiche, pois, lentille et vesce. D'autres foyers de domestication ont existé en Asie (riz) ou en Amérique du sud (maïs). La sélection intensive et la culture des phénotypes sélectionnés de ces espèces a amené aux variétés domestiques cultivées que nous connaissons aujourd'hui, exploitées sur des surfaces plus importantes et avec des rendements plus élevés (Stukenbrock and McDonald, 2008).

Comme tout organisme vivant, les plantes sont soumises en permanence à des attaques d'agents pathogènes. Néanmoins, la maladie reste une exception. La prévalence et les dommages engendrés par les agents pathogènes sont plus importants sur les espèces cultivées (Stukenbrock and McDonald, 2008). De ce fait les agents pathogènes représentent une menace sur l'agriculture justifiant les recherches qui leurs sont consacrés. Les plantes ont mis en place, au cours de leur évolution, des systèmes de défense reposant en partie sur des protagonistes moléculaires leur permettant de lutter contre ces agressions. Deux types de mécanismes peuvent être distingués :

- Les mécanismes passifs, qui correspondent à un processus constitutif, constant dans le temps, et qui impliquent à la fois des métabolites secondaires (composés phénoliques et alcaloïdes) et des défenses structurales (parois cellulaires, cuticules épidermiques et écorces).

- Les mécanismes actifs, qui sont induits lors du phénomène d'infection (ou de blessure) et sont basés sur la reconnaissance de l'agent pathogène par la plante hôte. Les principaux mécanismes sont la formation de papilles pariétales pour empêcher la pénétration tissulaire, la réaction

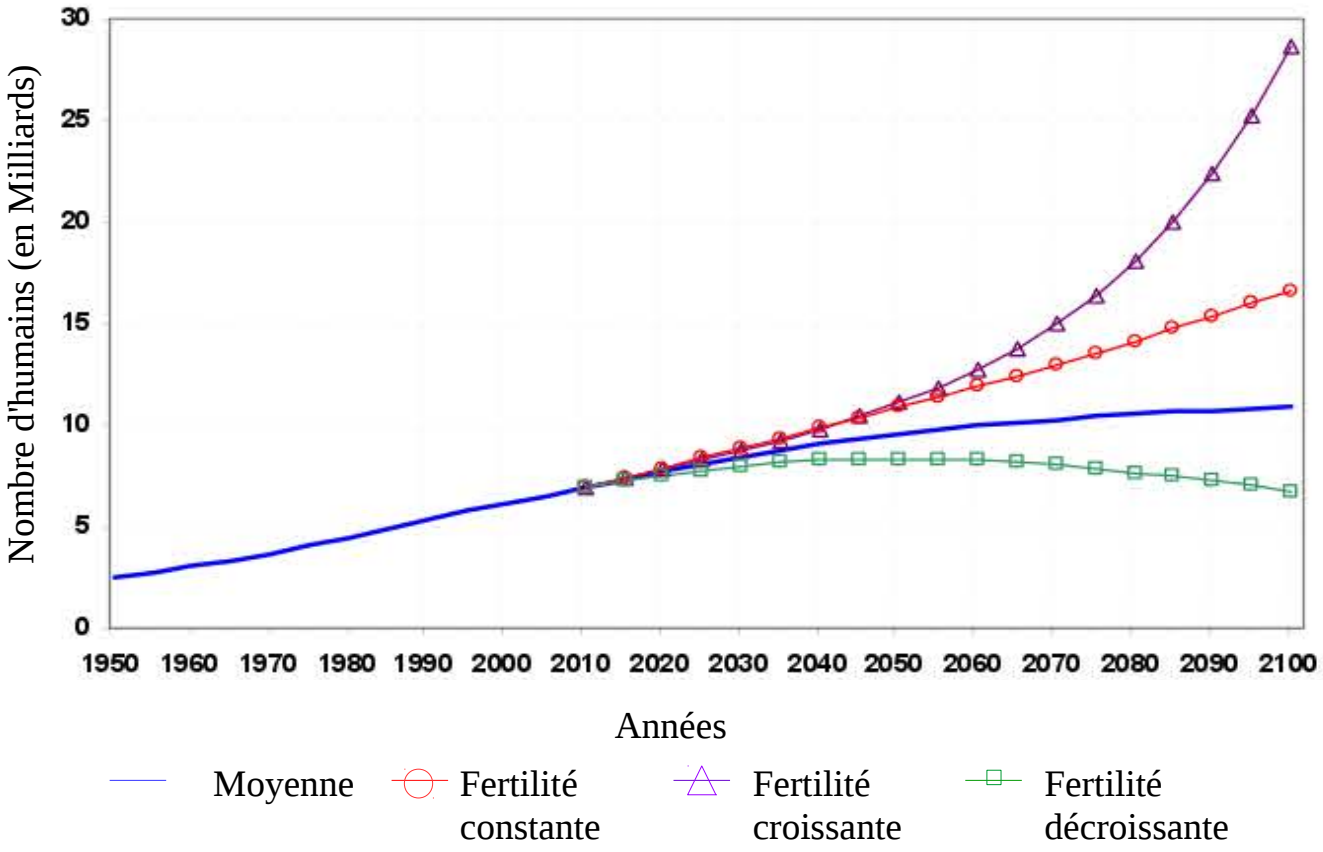


Figure 1 : Evolution de la population humaine mondiale entre 1950 et 2100 selon les taux de fertilité. D'après « Population Division of the Department of Econonmic and Social Affairs of the United Nations Secretariat (2013) » : World Population Prospects: The 2012 Revision. New York: United Nations

d'hypersensibilité caractérisée par une une mort cellulaire rapide au point d'infection et l'action des protéines PR (« Pathogenesis Related ») telles que des hydrolases ou des gluconases (Freeman and Beattie, 2008).

En terme d'évolution, l'interaction entre une plante et un pathogène présente l'intérêt de pouvoir étudier des phénomènes évolutifs sur des pas de temps courts. En effet, une interaction hôte-pathogène évolue constamment et rapidement, étant donné les importantes pressions de sélection réciproques. Ces dernières années, l'étude des interactions plantes-pathogènes est arrivée au stade de l'identification des déterminants moléculaires sous la forme des gènes et des protéines qu'ils codent. Mais la compréhension de l'évolution de ces interactions nécessite d'intégrer l'échelle moléculaire à des niveaux supérieurs, tels que l'individu et/ou la population. L'évolution de l'interaction hôte-pathogène peut être détectée par l'étude de la variation entre individus. Tout cela nécessite de s'intéresser à la variabilité inter-individuelle et à la structuration de cette diversité au sein et entre les populations.

1. Les interactions plante-parasite

La phytopathologie est la science qui étudie les maladies des plantes, notamment celles causées par les trois grands groupes de micro-organismes que sont les champignons, les oomycètes et les bactéries (Dodds and Rathjen, 2010). Un des modes de classification de ces organismes parasitaires est fonction de leurs stratégies d'infection. Dans ce cadre on distingue trois catégories :

-Les nécrotrophes, qui tuent les cellules de l'hôte (notamment par la sécrétion de toxines) et prolifèrent grâce aux nutriments ainsi relâchés. Ils présentent généralement une large gamme d'hôtes du fait de la généricité des mécanismes mis en œuvre.

-Les biotrophes, qui colonisent les cellules de leurs hôtes et s'en nourrissent sans entraîner la mort de cet hôte. L'interaction relève souvent d'une longue co-évolution au cours de laquelle le parasite s'est adapté à son hôte, et le parasite est ainsi spécialiste sur une gamme d'hôtes restreinte, bien qu'il existe des exceptions.

-Les hémibiotrophes, qui présentent un statut intermédiaire puisqu'ils démarrent généralement leurs cycles infectieux par une phase biotrophe avant une phase nécrotrophe.

Dans la nature, la maladie est une exception et la plupart des interactions plante-pathogène sont dites « non-hôte », car aucun des génotypes de la plante n'est un hôte pour le parasite. A l'inverse, lorsqu'au moins un des génotypes de la plante est infecté par le parasite, on parle de relation « hôte ». Dans ce cadre, l'interaction sera dite « incompatible » pour un des génotypes de la plante, s'il présente une résistance totale (ou qualitative) à un des biotypes du pathogène (celui-ci étant appelé avirulent dans ce cas). Au contraire, lors de la relation « compatible » la résistance de la plante peut s'exprimer (résistance quantitative) mais n'empêche pas le développement de la

maladie.

1.1. Les bases moléculaires de l'interaction

Cette partie s'intéresse aux différents modèles conceptuels des interactions plantes-pathogènes. Ces modèles ont progressivement pris en compte l'évolution de ces interactions, dénotant les progrès des connaissances sur les mécanismes moléculaires sous-jacents.

1.1.1. Le modèle gène-pour-gène, l'interaction statique

Le modèle central de reconnaissance hôte-pathogène en phytopathologie est celui de la relation gène-pour-gène (GPG) (Flor, 1971). Dans ce cadre, un allèle de résistance de la plante, codant un récepteur, reconnaît (ou non) un éliciteur produit par l'agent pathogène. Cette reconnaissance induit la mise en place des mécanismes de défense de type « réponse hypersensible » (ou HR abréviation de l'anglais « Hypersensitive Response ») conduisant à la résistance (R-Avr, ou R et Avr sont des traits dominants). Dans ce cas, on parle d'interaction incompatible , le pathogène ne se développe pas. Si la reconnaissance n'a pas lieu, la plante est sensible (soit parce qu'elle ne possède pas le gène de résistance adéquat : r-Avr, r-avr ; soit parce qu'il est contourné : R-avr). On parle alors d'interaction compatible, puisque le pathogène se développe (figure 2).

1.1.2. Le modèle en zigzag

Les avancées dans l'étude des interactions plante-micro-organisme ont permis d'établir un premier cadre conceptuel des interactions plantes-pathogènes prenant en compte la dynamique évolutive, modèle dit en zigzag (Jones and Dangl, 2006). Ce modèle, qui intègre le modèle GPG plus général, décrit les différents niveaux de la réponse immunitaire chez les plantes avec quatre stades distincts observés entre les deux protagonistes (figure 3), conduisant à la résistance ou à la sensibilité au pathogène. Au-delà, c'est un modèle de l'évolution des déterminants moléculaires de l'hôte et du pathogène impliqués dans l'interaction à plus long terme. Sans remettre en cause les phases de l'immunité décrites par ce modèle, des variantes centrées sur les aspects cellulaires et moléculaires sont apparues. Dodds and Rathjen, (2010) ont notamment proposé une variante basée sur les interactions plante-champignon, oomycètes et bactéries. En 2013, un modèle plus complet, comprenant en plus les interactions avec les insectes, les pucerons et les nématodes, permet de rediscuter le statut des effecteurs (voir section 1.1.2.1) (Dangl et al., 2013). L'intégration des virus, qui présentent des modes spécifiques d'interaction avec leurs hôtes, n'est toujours pas réalisée dans ces modèles.

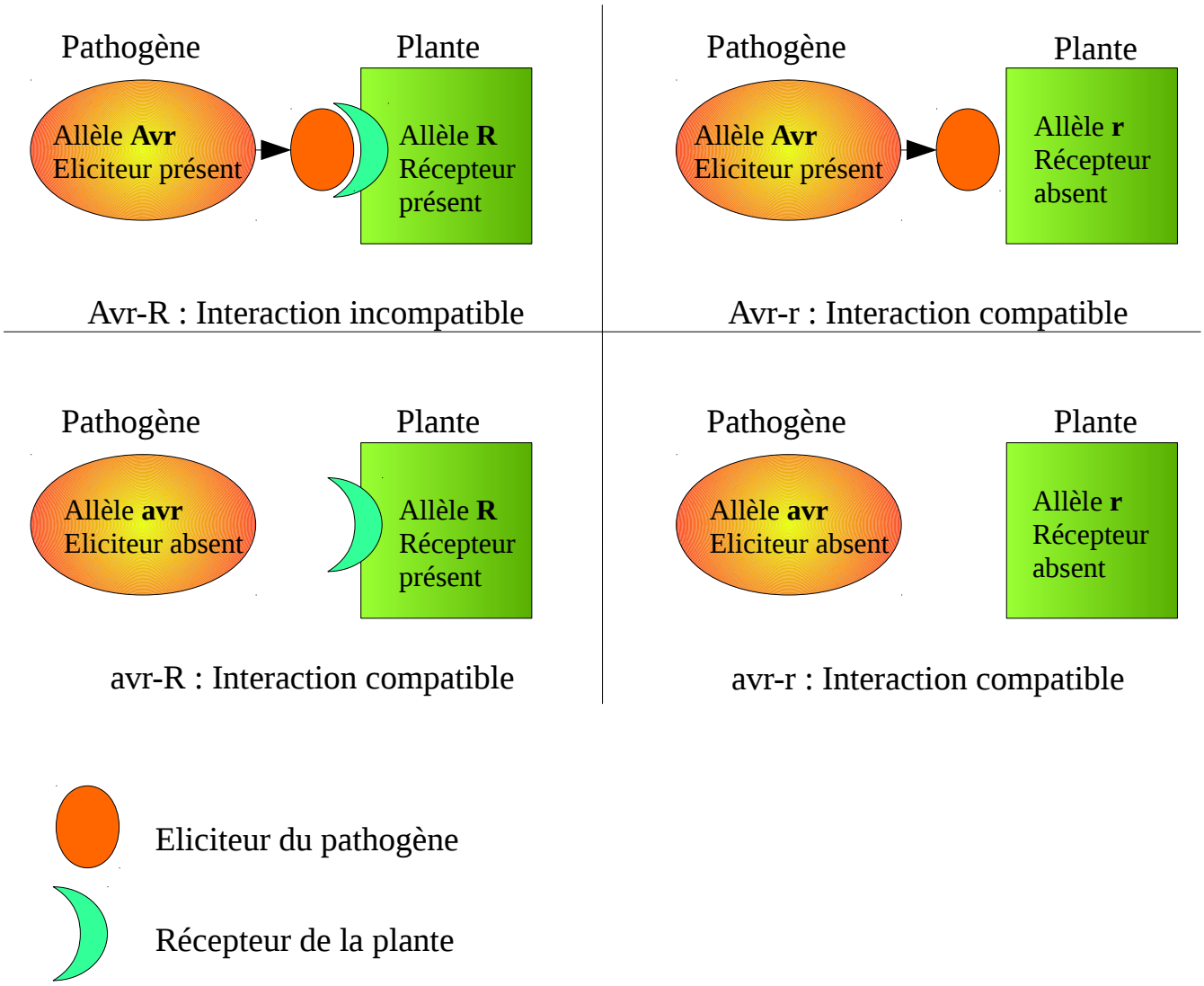


Figure 2 : Représentation schématique de la relation gène-pour-gène. (adapté de Mc Donald, 2004).
Dans le cas d'une interaction incompatible, le gène de résistance de la plante code un récepteur qui reconnaît un éliciteur produit par le pathogène. La reconnaissance de l'éliciteur du pathogène par le récepteur de la plante induit la mise en place des mécanismes de défense conduisant à la résistance (R-Avr). Dans tous les autres cas, la réaction est dite compatible et le pathogène se développe car, soit la reconnaissance n'a pas lieu (éliciteur absent, avr), soit (et/ou) la plante est sensible (r).

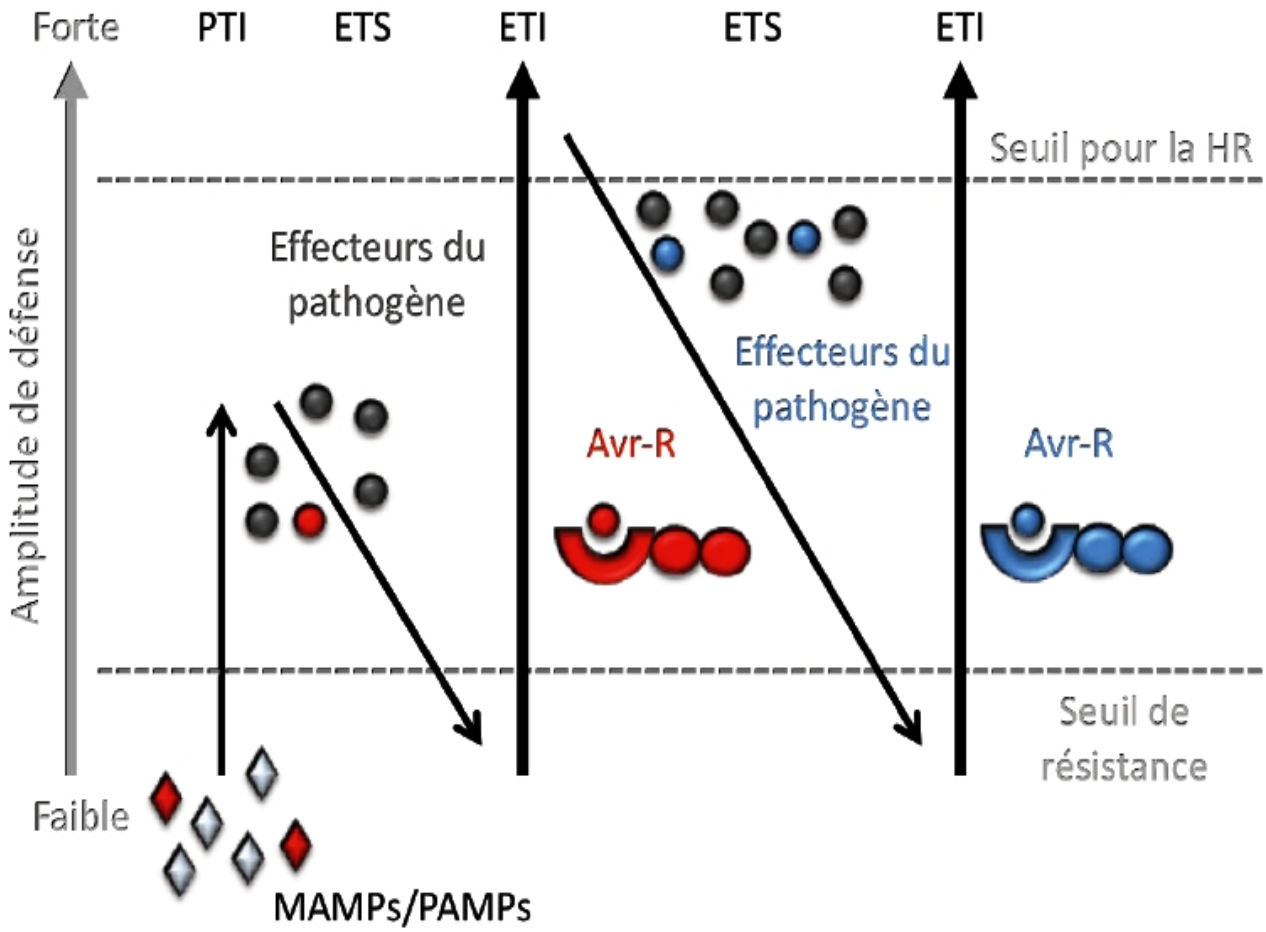


Figure 3 : Représentation schématique du modèle en zigzag (Jones and Dangl, 2006).

Phase 1 : La plante détecte les éliciteurs généraux (en rouge) du pathogène (Microbial/Pathogen Associated Molecular Patterns MAMPs/PAMPs) via ses récepteurs PRRs (Pathogen Recognition Receptors) déclenchant la première barrière de l'immunité (PAMP Triggered Immunity, PTI).

Phase 2 : Le pathogène virulent délivre des effecteurs qui interfèrent avec la première barrière de l'immunité (PTI) entrainant la sensibilité de la plante (Effector Triggered Susceptibility, ETS).

Phase 3 : Un effecteur du pathogène (en rouge) est reconnu par une protéine de résistance de la plante R (R-Avr), activant la deuxième barrière de l'immunité (Effector Triggered Immunity ETI) qui dépasse le seuil d'induction de la réponse hypersensible (Hypersensitive Response, HR) et induit la mort cellulaire, stoppant la croissance du pathogène.

Phase 4 : Des isolats du pathogène ont perdu l'effecteur rouge et/ou accumulé de nouveaux effecteurs (en bleu) pouvant de nouveau interférer avec le système immunitaire de l'hôte (ETS). Au cours de l'évolution, la sélection favorise l'apparition de nouveaux allèles R chez la plante qui reconnaissent les effecteurs du pathogène (ETI).

1.1.2.1. Mécanismes de défense non-spécifiques

Dans ces modèles, le premier stade de l'interaction d'un point de vue évolutif correspond à la réponse immunitaire primaire ou basale. Des éliciteurs généraux (MAMPs/PAMPs, « Microbe/Pathogen Associated Molecular Patterns ») sont libérés par le pathogène et sont reconnus par des récepteurs de la plante (PRR, « Pattern Recognition Receptors »), déclenchant ainsi les mécanismes de défense dits « non-hôtes » ou « généraux » (PTI, « PAMP-Triggered Immunity »). Ces PRR appartiennent à la famille des receptor-like kinases (RLK) ou des receptor-like protein (RLP) se différenciant par la présence ou l'absence d'un domaine kinase. Actuellement plus d'une dizaine de couple éliciteurs/récepteurs ont été identifiés, essentiellement entre des PRR d'*Arabidopsis thaliana* ou du riz et des éliciteurs bactériens (Macho and Zipfel, 2014). Concernant les champignons pathogènes, les gènes de type CERK1 reconnaissent des oligomères de chitine libérés par la dégradation des parois cellulaires de champignons pathogènes, induisant des mécanismes de défense générique (Liu et al., 2012; Miya et al., 2007; Wan et al., 2008).

1.1.2.2. Les effecteurs, molécules clefs de l'interaction

Un pathogène pourra contourner la défense primaire de type PTI s'il est capable de libérer des effecteurs, qui vont interagir avec la cellule hôte et interférer avec cette première barrière de l'immunité, rendant ainsi la plante sensible. Cette deuxième phase du modèle en zigzag est appelée ETS (« Effector-Triggered Susceptibility »). Au cours de celle-ci, des effecteurs peuvent interférer avec la réponse PTI, conduisant à la sensibilité. Les avancées récentes en terme de séquençage haut-débit des génomes (Mardis, 2008) ont permis de mettre à jour de larges répertoires de protéines sécrétées qui sont considérées comme effecteurs putatifs (Raffaele and Kamoun, 2012). Le rôle de ces effecteurs dans les interactions plantes-parasites s'est ainsi élargi vers un rôle de répresseur du PTI et d'activateur de l'ETS, comme représenté dans le modèle de Dangl et al., (2013) (figure 4).

La question de la nature des effecteurs n'est ainsi pas encore réglée. Un effecteur est généralement présenté comme une protéine sécrétée par un agent pathogène vers la cellule de l'hôte et qui à pour but d'interférer avec son immunité. Or, une grande variété de protéines sont désormais incluses dans les effecteurs avec pour rôle de promouvoir la croissance des agents pathogènes biotrophes (effecteurs apoplastique et cytoplasmique) ou nécrotrophes (enzymes dégradant la paroi cellulaire et toxines spécifique de l'hôte). Ces effecteurs peuvent être des protéines, des peptidoglycanes, des polysaccharide, ou encore des peptides non ribosomaux (Schneider and Collmer, 2010). Les recherches d'effecteurs cytoplasmiques menées sur les génomes séquencés de *Phytophthora infestans*, *P. sojae* et *Hyaloperonospora arabidopsis* ont permis de trouver des motifs d'acides aminés conservés, de type RXLR, LXFLAK ou CHXC, qui sont impliqués dans la translocation des protéines au sein des cellules de la plante (Baxter et al., 2010;

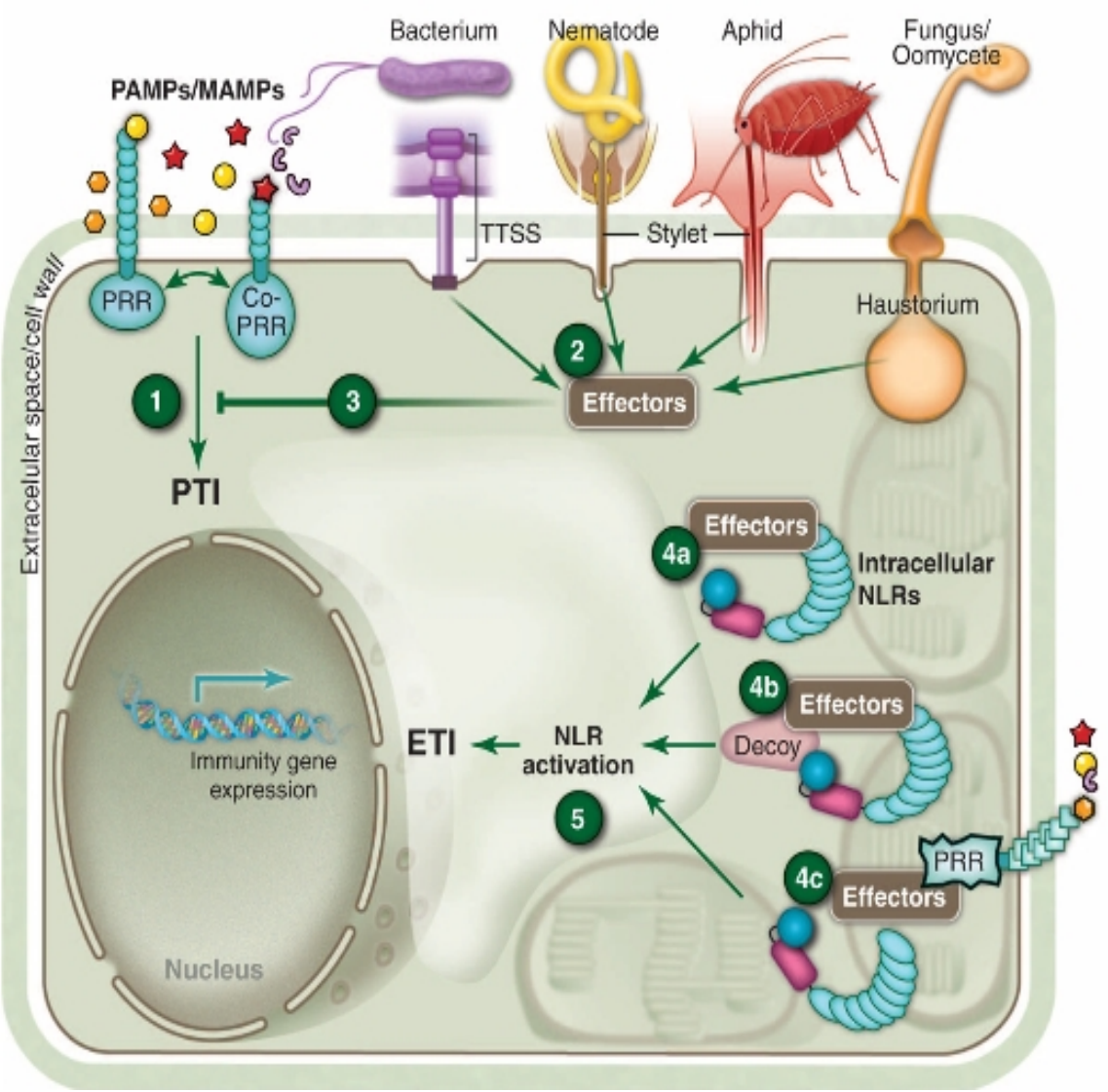


Figure 4 : Représentation schématique du système immunitaire des plantes

Les agents pathogènes des différentes classes expriment des PAMP et MAMP lorsqu'ils colonisent les plantes (les couleurs sont codées selon les agents pathogènes). Les plantes les perçoivent via les PRR extracellulaires et initient l'immunité (PTI; étape 1). Les pathogènes délivrent alors des effecteurs de la virulence, à la fois dans l'apoplaste des cellules végétales et à l'intérieur des cellules de la plante (étape 2). Ces effecteurs sont adressés à des emplacements subcellulaires spécifiques où ils peuvent réprimer la PTI et permettre l'établissement de la virulence (étape 3). Des récepteurs NLR intracellulaires peuvent percevoir les effecteurs de différentes manières : par l'interaction directe entre l'effecteur et le récepteur (étape 4a); par la détection de l'interaction entre l'effecteur et une protéine leurre qui mime structurellement une cible de l'effecteur (étape 4b) ; ou en détectant une modification d'une cible de l'effecteur dans la cellule hôte comme le domaine cytosolique d'une PRR (étape 4c). D'après Dangl et al., (2013)

Tyler et al., 2006; Haas et al., 2009). Des répertoires d'effecteurs candidats avec plusieurs centaines de membres ont pu être identifiés grâce à la recherche de séquences basée sur ces motifs (Raffaele et al., 2010).

Concernant les champignons phytopathogènes, la tâche d'identification d'effecteurs s'avère difficile du fait de l'absence de motifs conservés parmi les effecteurs décrits (Rafiqi et al., 2012). Cette absence peut s'expliquer en partie par la spécialisation de ces gènes qui sont souvent spécifiques à une espèce ou à un genre donné (Duplessis et al., 2011a). Ainsi, la recherche systématique de gènes codant pour des effecteurs candidats chez les champignons phytopathogènes s'est basée sur des critères plus larges : signal de sécrétion, petites protéines, richesse en cystéine et signal de sélection positive (Presti et al., 2015; Saunders et al., 2012; Stergiopoulos and de Wit, 2009). Une partie de ces critères est encore en discussion au sein de la communauté . Les effecteurs ont été classiquement considérés comme devant être des petites protéines (moins de 300 acides aminés) (Hacquard et al., 2012). Mais la découverte d'AvrM-A de *Melampsora lini*, un effecteur de 343 acides aminés, a remis en cause cette limite (Ve et al., 2013). Par ailleurs, la recherche des signaux de sécrétion dans les gènes codant ces protéines est complètement dépendante de la richesse des bases de données (Lo Presti et al., 2015). Le risque de faux négatifs est donc élevé.

1.1.2.3. La relation R-Avr

Dans plusieurs pathosystèmes, il a été montré que la plante est capable de réagir au contournement de sa défense primaire par la reconnaissance de l'un des effecteurs du pathogène. Cette reconnaissance est médiée le plus souvent par une protéine de résistance (récepteur intracellulaire de type Nucleotide Binding domain-Leucine-Rich Repeat, NB-LRR). Elle conduit alors à des réactions de défense de la plante (Effector-Triggered Immunity) dont l'intensité est plus importante que pour la réponse primaire de type PTI. Ces réactions conduisent à une HR et définissent un second niveau de reconnaissance dit hôte-spécifique. Celui-ci, basé sur l'interaction R-Avr (relation gène-pour-gène), correspond à la réponse immunitaire secondaire de la plante. C'est la relation entre le facteur d'avirulence du parasite et de la protéine de résistance de la plante décrite par le modèle GPG de Flor (Flor, 1971) (figure 2).

Concernant les champignons, plusieurs couples R-Avr on pu être mis à jour dans plusieurs pathosystèmes et leur implication dans la mise en place d'une HR a été démontrée. A titre d'exemple chez l'agent de la rouille du lin *M. lini* qui est un modèle d'étude pour les interactions plante-agent des rouilles et qui a notamment été au cœur du modèle GPG décrit par Flor, plusieurs effecteurs correspondant à des gènes d'avirulence ont pu être décrit. Les facteurs d'avirulence AvrM, AvrL567, AvrP4 et AvrP123 entraînent une HR quand ils sont exprimés dans des plantes exprimant les protéines de résistance de type NB-LRR correspondantes (M, L5/L6/L7, P4 et P/P1/P2/P3 ; Ellis et

al., 2007). De plus, une localisation de ces protéines dans les cellules de l'hôte a pu être montré pour certaines de ces protéines de *M. lini* (Rafiqi et al., 2010). Pour d'autres l'interaction avec des protéines intracellulaires de type NB-LRR indique la détection des protéines fongiques à l'intérieur des cellules de l'hôte. Cette localisation confère à ces protéines fongiques le caractère d'effecteur. Toutefois, leur rôle primaire dans l'infection n'est toujours pas déterminé (Petre et al., 2014).

1.2. Les mécanismes évolutifs

Chacun de ces modèles d'interaction entre la plante et son/ses parasites engendrent une très forte pression de sélection sur les deux partenaires. Des modèles de théorie générale de coévolution des gènes du pathogène et son hôte ont été développés pour décrire u
n cadre conceptuel de cette évolution. Les deux modèles de coévolution dynamique principalement utilisés sont la guerre des tranchées et la course aux armements (figure 5).

La guerre des tranchées décrit l'évolution des reconnaissances hôte-pathogène, dans laquelle deux ou plusieurs allèles de gènes effecteurs coexistent dans la population du pathogène. Cette théorie repose sur la sélection balancée, dépendante de la fréquence des allèles. L'allèle offrant la plus grande contribution à la valeur sélective (« fitness ») des agents pathogènes, augmente en fréquence. À son tour, l'allèle de l'hôte correspondant augmente en fréquence dans la population. Ce processus aboutit à la réduction de la valeur sélective et donc de la fréquence de l'allèle du gène effecteur, qui est ensuite suivie par une diminution de la fréquence de l'allèle hôte correspondant et ainsi de suite. Ce modèle est gouverné par de la sélection négativement dépendante de la fréquence allélique, et les fréquences alléliques des gènes correspondant à l'agent pathogène et hôte oscillent au fil du temps. Ce modèle suppose que la possibilité d'innovation (génération de nouveaux allèles) est limitée et/ou qu'elle a un coût, et c'est l'évolution des fréquences d'allèles existants qui engendre les cycles évolutifs de l'interaction hôte-pathogène (Stahl et al., 1999).

Le modèle alternatif est celui de la course aux armements. Il décrit l'interaction hôte-pathogène comme allant vers une innovation perpétuelle des gènes d'avirulence d'un côté et de résistance de l'autre, générant un équilibre dynamique entre les deux protagonistes. La fréquence de l'allèle offrant la meilleure fitness au pathogène augmente rapidement en fréquence dans la population de pathogènes, et remplace finalement les allèles moins adaptés. A son tour, un nouvel allèle d'un gène de résistance de l'hôte, lui permettant de se soustraire à l'effet de l'effecteur du pathogène, apparaît, et sa fréquence allélique augmente rapidement pour finalement se fixer dans la population. Ce cycle se répète indéfiniment. Génétiquement, les gènes codant les protéines impliquées présentent une évolution rapide engendrant une plus grande quantité de remplacements d'acides aminés entre les espèces. À chaque épisode sélectif, un balayage sélectif réduit la diversité

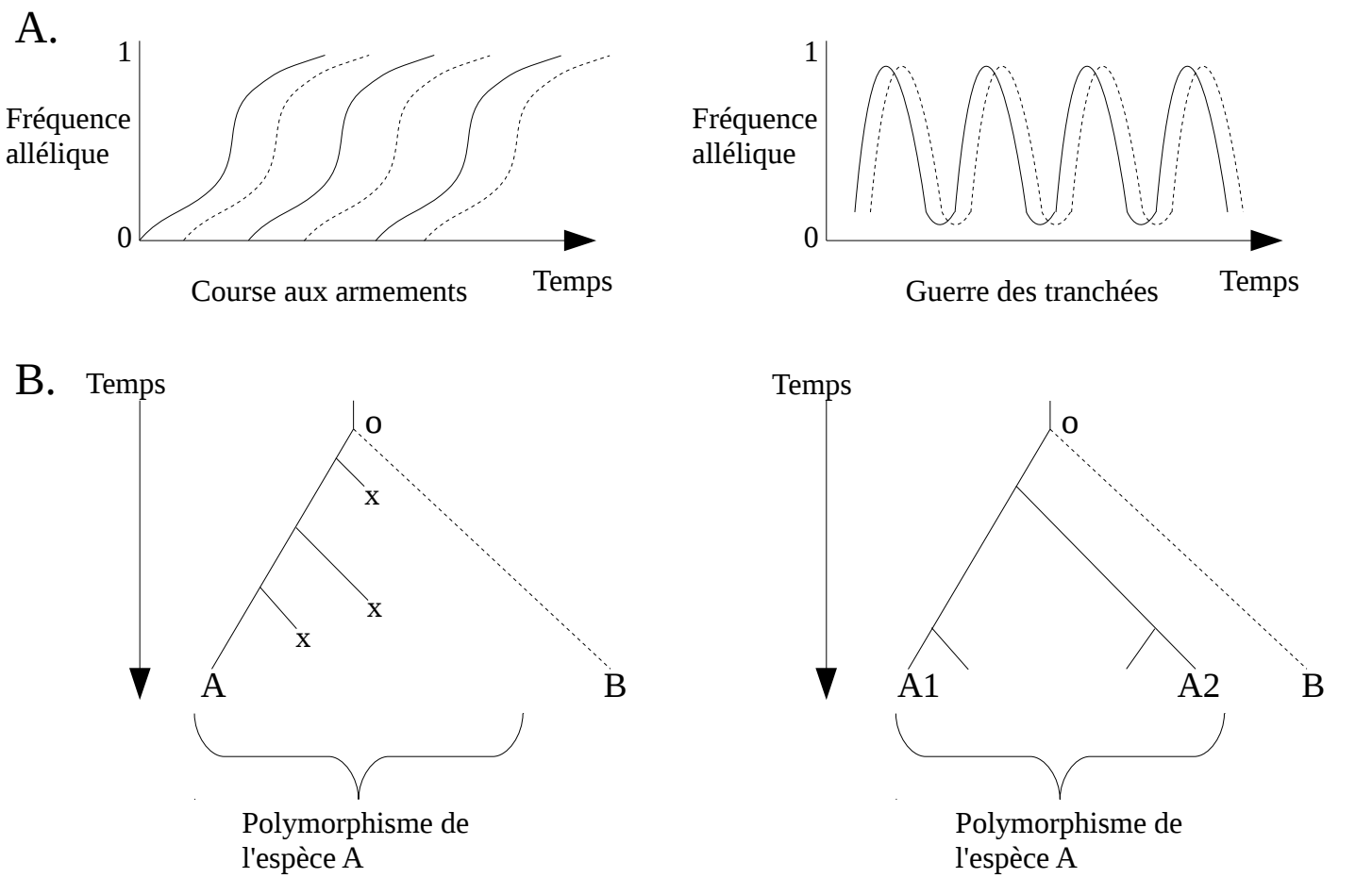


Figure 5 : La course aux armements et la guerre des tranchées, les deux scénarios d'évolution des interactions hôte-pathogène et leurs effets sur les polymorphismes de l'ADN et la divergence.

A. Les variations de fréquences alléliques dans le modèle de la course aux armements (à gauche) et de la guerre des tranchées (à droite). La ligne continue et la ligne pointillée correspondent respectivement à la fréquence allélique de l'effecteur du pathogène et du gène de résistance de l'hôte. B. Généalogie d'un locus impliqué dans la course aux armements (à gauche) ou la guerre des tranchées (à droite). Les lignes pleines indiquent la généalogie au sein d'une espèce, alors que les lignes pointillées sont pour une espèce voisine. Dans le scénario de la course aux armements, un seul allèle (indiqué par A) est choisi quand il confère une fitness supérieure. Les autres allèles disparaissent de la population (indiqué par 'x'). Dans le scénario de la guerre des tranchées, deux ou plusieurs allèles (indiqués comme A1 et A2) sont maintenus dans l'espèce pour une longue période. C. Signatures typiques résultant de la course aux armements (à gauche) et la guerre des tranchées (droite). Adapté de Terauchi and Yoshida (2010).

génétique et donc le polymorphisme intra-espèce (Terauchi and Yoshida, 2010).

Ainsi, ces théories décrivent un équilibre dynamique entre les deux acteurs et la répétition des cycles (résistance puis sensibilité chez la plante et avirulence puis virulence chez l'agent pathogène). Côté pathogène, le passage d'avirulence à virulence (dans le modèle gène-pour-gène) inclut plusieurs types de modifications sur le plan génétique (allant de la mutation ponctuelle à la délétion/insertion d'une région chromosomique) (Gout et al., 2006; Khang et al., 2008). L'essentiel étant pour être sélectionné que le changement confère une perte de reconnaissance par la plante hôte et génère ainsi une interaction compatible.

2. Analyse du polymorphisme

Les épisodes adaptatifs laissent des signatures spécifiques au sein du polymorphisme génétique au voisinage génomique de la région où se trouvait l'allèle favorable. La stratégie utilisée pour identifier ces régions part d'un principe simple : prédire les caractéristiques attendues du patron de polymorphisme sous une hypothèse de neutralité (l'hypothèse nulle au sens statistique), et vérifier la conformité des données observées le long du génome.

Cette stratégie, simple de prime abord, nécessite néanmoins de connaître la structure populationnelle des données, d'en connaître les facteurs de structuration et les caractéristiques, avant de pouvoir détecter les évènements de sélection. En effet, l'hypothèse de neutralité doit être formulée sous la forme d'un modèle démographique correspondant à la population observée.

2.1. Modèles de génétique des populations

2.1.1. Notion de population

Le principe de la génétique des populations est l'étude des fréquences alléliques portées par les individus au niveau de marqueurs génétiques distribués dans le génome, dans le but de comprendre la biologie des populations, la spéciation et/ou l'évolution moléculaire. Cette discipline scientifique s'intéresse aux caractéristiques de ces fréquences d'allèles au sein et entre les populations d'individus sous forme de la variabilité génétique et de son partitionnement. Ces fréquences alléliques sont en effet influencées par les caractéristiques des populations (structure spatiale, flux géniques, histoire démographique) (Wright, 1931).

Le terme 'population' est très utilisé en biologie mais reste difficile à définir. Il est généralement défini comme un groupe d'individus appartenant à la même espèce, localisé au même endroit et au même moment et capables de se reproduire entre eux. A partir d'un échantillon d'individus, la génétique des populations peut permettre de définir des populations *a posteriori*, de caractériser ces populations et enfin d'inférer les paramètres démographiques et/ou historiques.

C’est dans ce cadre que de nombreux indices ont été définis sous deux catégories : les indices intra-populationnels (décrire la répartition de la variabilité génétique des individus au sein d’une population) et les indices inter-populationnels (décrire la répartition de la variabilité génétique des populations entre elles). Ces différents indices sont résumés dans l’encadré 1 Ils vont permettre principalement deux choses. Tout d’abord, une vision descriptive de l’échantillon permettant d’identifier son polymorphisme. Ensuite, une vision populationnelle quantifiant la part de la variance génétique expliquée par la structure populationnelle (souvent extrapolée comme un indice de différenciation génétique entre populations) (Halkett et al., 2005).

2.1.2. Les facteurs d'évolution des populations

L’évolution agit sur les populations via cinq forces évolutives : la dérive, la sélection, la migration, la mutation et la recombinaison. Nous n'aborderons ici que les effets de ces forces au niveau populationnel, car leurs effets moléculaires peuvent être assez différents. Ces cinq forces peuvent être définies de la manière suivante :

La **dérive génétique** correspond à la variation aléatoire des fréquences alléliques au sein d’une population de taille finie au cours des générations. Cette force va donc fixer, de manière aléatoire, des allèles neutres voire légèrement délétères dans les populations, mais la plupart des allèles créés par la mutation ou apportés par la migration seront éliminés au fil des générations car ils sont initialement en fréquence faible. La dérive conduit donc à la perte de variations génétiques au sein des populations.

La **sélection** correspond à l’ensemble des phénomènes qui augmentent ou diminuent la fréquence d'un phénotype en raison de l’adaptation des individus à leurs environnements. Il existe trois types de sélection sur le phénotype :

La sélection stabilisante maintient les phénotypes optimaux au détriment de la diversité. Cette force favorise un optimum unitaire au sein de la population. Son effet le plus important est donc d’éliminer les génotypes nouveaux apparaissant par mutation ou migration. Ainsi, dans un environnement stable, une population ayant atteint un haut niveau d'adaptation maintient les génotypes les plus adaptés au fil des générations.

La sélection directionnelle favorise également un optimum au sein de la population, mais en générant un décalage systématique des fréquences alléliques dans une certaine direction. Cette force est effective, chez des populations évoluant dans un environnement se modifiant de manière constante, et aboutit à un nouvel état d'adaptation.

Encadré 1: Paramètres de génétique des populations

La génétique des populations a pour but d'étudier la variabilité génétique et son partitionnement entre individus et entre populations. De ce fait, les paramètres calculables peuvent être divisés en deux catégories selon le niveau d'organisation considéré : les paramètres intra-populationnels qui reflètent la diversité génétique au sens large, et les paramètres inter-populationnels qui permettent de quantifier les différences de fréquences d'allèles entre populations.

Les paramètres intra-populationnels

N : Correspond à la taille de la population (nombre d'individus). A ne pas confondre avec *Ne* (taille efficace de la population).

G/N : Correspond au nombre de génotype ramené au nombre d'individu par population (correspond donc à la richesse en génotype d'une population).

Ho (observed heterozygosity) : Correspond à l'hétérozygotie observée au sein d'une population. Elle est calculée à partir de la fréquence des individus hétérozygotes (nombre d'individus hétérozygotes divisé par la taille de la population ***N***).

He (expected heterozygosity) : Correspond à la fréquence théorique des hétérozygotes d'une population **(**diversité génétique de Nei [1973]). Elle est calculée à partir des fréquences alléliques. S'il y a n allèles en fréquences f1, f2, f3, ...fn, la fréquence théorique des hétérozygotes sera

$$He = \left(N / (N-1) \right) \times 1 / L \sum_{L} \left(1 - \sum_{n} f_n^{\,2} \right)$$

avec *L* = nombre de loci.

F_{IS} **:** Nommé aussi indice de fixation ou coefficient de consanguinité est calculé selon la formule :

$\boldsymbol{F_{IS} = (he - ho)/ he = 1 - (ho/he)}$

Il reflète la différenciation des individus à l'intérieur des populations (F_{IS} = 1 signifie fixation complète (cas d'autofécondation), F_{IS} inférieur à 1 : hétérozygotie excédentaire, F_{IS} = 0 : population en équilibre de Hardy-Weinberg).

Les paramètres inter-populationnels

$\boldsymbol{F_{ST}}$ (indice de fixation) : Correspond à un paramètre de différenciation. Il est défini comme la répartition des fréquences alléliques entre les subdivisions d'une population. Il représente la corrélation entre les allèles à l'intérieur d'une sous-population par rapport à l'ensemble des sous populations. La différenciation des populations par rapport au total (F_{ST}) est calculée en fonction des paramètres F_{IS} (différenciation des individus à l'intérieur des populations) et F_{IT} (différenciation des individus par rapport au total). Ils sont liés par la relation :

$$\boldsymbol{F_{ST} = 1 - \frac{(1 - F_{IT})}{(1 - F_{IS})}}$$

Si $0 < F_{ST} < 0{,}05$: différenciation faible
$0{,}05 < F_{ST} < 0{,}15$: différenciation modérée
$0{,}15 < F_{ST} < 0{,}25$: différenciation importante
$F_{ST} > 0{,}25$: différenciation très importante

Ces différents paramètres peuvent mesurer les forces évolutives régissant les populations.

Ainsi deux populations présentant $F_{IS} > 0$ (donc étant en excès d'homozygotes) peuvent être différenciées au moyen du *G/N* et du *LD* : un *G/N*<1 et un fort déséquilibre de liaison sont le signe d'une population clonale alors qu'une population présentant un *G/N*=1 et pas de déséquilibre de liaison est le signe d'une population panmictique présentant une reproduction sexuée (Halkett et al., 2005).

Le déséquilibre de liaison DL correspond à l'absence d'indépendance statistique entre les génotypes de deux locus. Ces deux locus seront dits en déséquilibre de liaison ou liés car ils apparaissent ensemble, au sein des individus d'une population, plus fréquemment qu'au hasard.

La sélection diversifiante favorise plus d'un optimum au sein de la population. Cette force s'applique à les populations présentes dans un environnement hétérogène. Plusieurs génotypes sont favorisés au détriment des intermédiaires.

La **migration** consiste en des échanges génétiques entre populations. Le taux de migration (nombre d'individus qui migrent) n'étant pas forcément égale entre la population dite donneuse et la receveuse. En terme de diversité génétique, cette force présente une tendance homogénéisatrice entre populations mais elle diminue l'effet de la dérive au sein des populations

La **mutation** engendre, de manière aléatoire, de la diversité génétique au sein des individus et donc des populations. Il s'agit de la source ultime de variation génétique. En particulier, le taux de mutation n'est pas constant au sein du génome, engendrant des zones riches en mutations appelées points chauds mutationnel (Rogozin and Pavlov, 2003).

La **recombinaison** est le phénomène conduisant à l'apparition, dans un individu, de gènes dans une association différente de celles observées chez les individus parentaux. Ces nouvelles combinaisons de génotypes pouvant résulter en des capacités adaptatives différentes. Cette force tend également à augmenter la diversité génotypique.

2.1.3. Les assignations Bayésiennes

Une partie de la génétique des populations est dédiée à l'étude du partitionnement local des individus permettant d'accéder aux groupes ancestraux. La non prise en compte du partitionnement local engendre des biais dans les analyses génétiques (Whitlock and McCauley, 1999) si la migration n'est pas prise en compte. C'est dans ce cadre que sont apparues les méthodes d'assignations Bayésiennes qui permettent de regrouper *a posteriori* des individus dans des populations génétiquement homogènes, sur la seule base des données génotypiques (Beaumont and Rannala, 2004).

Ces méthodes sont basées sur un des modèles fondateurs de la génétique des populations à savoir le modèle de Hardy-Weinberg. Il permet, sous certaines conditions de mesurer les fréquences génotypiques à partir des fréquences alléliques. Les conditions de ce modèle sont :

- La population est panmictique (les couples se forment au hasard (panmixie), et leurs gamètes se rencontrent au hasard (pangamie)). Tout se passe comme si tous les gamètes étaient mis en commun.

- La population est considérée comme infinie (elle est en réalité suffisamment grande pour s'affranchir des biais d'échantillonnage).

- Il ne doit y avoir ni sélection, ni mutation, ni migration (pas de perte/gain d'allèle).

- Les générations successives sont discrètes (pas de croisement entre générations différentes).

Dans ces conditions, la diversité génétique de la population se maintient et doit tendre vers un équilibre stable de la distribution génotypique (Luikart et al., 2003; Graffelman et al., 2013).

Les assignations Bayésiennes sont basées sur un modèle décrivant la structure du polymorphisme génétique en fonction de la structuration des individus en populations dont la vraisemblance est estimée grâce à des méthodes MCMC (pour Markov Chain Monte Carlo) qui sont une classe de méthodes d'échantillonnage se basant sur l'exploration aléatoire de l'espace des paramètres (chaînes de Markov = exploration sans mémoire, Monte Carlo = au hasard). Elles consistent à minimiser, au sein de la population (donc entre les individus la composant), les déséquilibres de liaisons et le déséquilibre de Hardy-Weinberg. Ceci permet donc d'avoir accès à des groupes ancestraux sur lesquels il devient possible de mesurer les indices sans les effets confondants de la sous structuration (voir encadré 1).

2.1.4. La coalescence

La coalescence est une manière de modéliser l'effet des différentes forces évolutives sur la généalogie d'individus appartenant à un échantillon. Elle est un puissant outil conceptuel pour la génétique des populations, permettant d'examiner l'effet de différents paramètres sur la structuration des populations et leur polymorphisme. Un large éventail de phénomènes biologiques peut être modélisé en utilisant cette approche (Beaumont and Rannala, 2004). Le modèle le plus simple (et servant souvent d'hypothèse nulle) est basé sur le scénario d'évolution décrit dans le modèle de Wright-Fisher. L'évolution des séquences sera décrite selon les termes de l'hypothèse neutraliste de l'évolution moléculaire de Kimura (Kimura, 1983a). C'est un modèle répartissant les individus en populations où la dérive et la reproduction sont les seules forces évolutives en présence. Il repose sur des populations à l'équilibre et ne subissant pas de sélection ni de migration, présentant un nombre N d'individus (ainsi $2N$ copies alléliques dans la population pour des individus diploïdes). Dans ce modèle la taille des populations est constante, elles ne subissent pas de subdivision et se reproduisent en panmixie stricte. Les conséquences des propriétés du modèle sont qu'un gène peut être transmis en plusieurs copies à la génération suivante ou, au contraire, ne pas être transmis du tout (il modélise ainsi le processus de dérive).

L'évolution des marqueurs moléculaires est décrite par un modèle neutre correspondant à l'hypothèse neutraliste de Kimura. Les mutations favorables sont rares, les mutations délétères sont été éliminées par sélection naturelle très rapidement et seules restent les mutations sélectivement neutres. Seules ces mutations sont considérées dans le modèle de coalescence. Chaque mutation apparaît sur un nouveau site de la séquence puisqu'on considère un nombre infini de sites (Kimura, 1983b). Les mutations apparaissent donc au hasard au fil des générations et augmentent en fréquence, se fixent (rarement) ou disparaissent (souvent) selon la seule force de la dérive. Au sein

de l'échantillon considéré, il n'est pas possible d'observer toutes les mutations apparues dans la population mais uniquement celles apparues dans l'histoire de cet échantillon (dans tous les ancêtres des individus échantillonnés jusqu'à l'ancêtre commun le plus proche). Ce modèle peut être étendu pour relâcher certaines hypothèses (reproduction sexuée, population de taille constante, présence de sous-populations et flux de gènes).

La coalescence peut être utilisée comme outil analytique (on calcule les probabilités de jeux de données selon les valeurs des paramètres) ou comme outil de simulation. Pris dans le sens d'outil de simulation, elle permet de réaliser un grand nombre de simulations et de générer un jeu de données à chaque fois (pour regarder la distribution des indices en fonction des valeurs de paramètres). La construction des arbres de coalescence nécessite l'utilisation de marqueurs neutres et suffisamment d'échantillons (individus) pour que le nombre de combinaisons soit totalement indépendant du type allélique (Estoup and Guillemaud, 2010). Les algorithmes procèdent à rebours, en partant des échantillons et en modélisant les évènements de coalescence (fusion de lignées) en remontant dans le temps jusqu'à l'ancêtre commun le plus récent entre tous les individus étudiées (nommé MRCA pour « Most Recent Common Ancestor »).

2.1.5. L'inférence par l'approche Bayésienne approchée

On peut également estimer des paramètres évolutifs avec des méthodes spécifiques, telle que l'ABC (« Approximate Bayesian Computation ») qui utilise les algorithmes de simulation de données décrits précédemment et des statistiques résumées (ex: *Ar*, *He*, F_{ST},…) pour approcher la vraisemblance d'un échantillon par un algorithme d'acceptation/rejet. La vraisemblance est difficile à estimer car il faut intégrer sur toutes les généalogies possibles. L'ABC permet de contourner le problème en s'appuyant sur des simulations. Les distributions *a posteriori* des paramètres de chacun des modèles sont déterminées en remplaçant le calcul de la vraisemblance (probabilité des données observées à partir des valeurs des paramètres du modèle) par une mesure de similarité entre les données observées et simulées. Ces simulations peuvent également servir à comparer différents scénarios de coalescence entre eux et déterminer le plus vraisemblable (celui qui génère des valeurs de statistiques résumées les plus proches de celles calculées directement avec le jeu de données). Ainsi, la méthode va calculer quel est le scénario le plus vraisemblable parmi ceux qu'on lui propose. Une des limites est que le programme identifiera toujours un scénario comme étant le meilleur, même si aucun de ceux proposés ne s'approche de la réalité de la généalogie des échantillons.

2.2. Détecter la sélection, un défi historique

Lorsque les populations ont été identifiées, il devient possible d'analyser l'effet de la sélection naturelle à l'échelle des populations. La stratégie employée part d'un principe simple : prédire les caractéristiques du patron de polymorphisme, sous l'hypothèse de neutralité (autrement appelée hypothèse nulle), et vérifier si les données observées s'y conforment. C'est dans ce contexte que les études de génomique des populations prennent tout leur sens, comme elles s'intéressent à l'ensemble des données du génome d'un lot d'individus, elles permettent de discriminer la sélection de la dérive et ainsi potentiellement de trouver des gènes d'intérêt impliqués dans l'adaptation d'un organisme à son environnement. Auparavant la détection des régions impliquées dans les épisodes de sélection était difficile à cause de la faible densité en marqueurs moléculaires disponibles. Le développement des techniques de séquençage aide à lever ce verrou.

2.2.1. L'essor de la génomique

2.2.1.1. L'évolution des techniques de séquençage

La technique de séquençage mise au point par Frederick Sanger, en 1977, a permis l'accès au potentiel énorme de l'étude des génomes. De nombreuses améliorations (synthèse chimique automatisée des amorces, électrophorèse capillaire) ont permis à cette technique d'atteindre aujourd'hui des longueurs de séquences de près de 1000 paires de bases de façon automatisée, pour 384 échantillons par run, et 24 runs en une journée, soit 10 mégabases (Mb) par jour. Malgré cela, cette technique de séquençage représente un coût très élevé pour un génome (1500 $ par Mb, soit cinq millions de dollars pour un génome humain).

Au cours de la dernière décennie, de nouvelles techniques sont apparues, reposant sur des principes différents de ceux de la méthode de Sanger. Ces types de séquençage présentent un intérêt économique considérable par rapport au séquençage traditionnel. Par exemple, pour une même couverture de génome humain, le coût d'un séquençage nouvelle génération réalisé avec les automates 454 ou Illumina est estimé à 70 fois moins que celui d'un séquençage Sanger (Glenn, 2011). Parmi ces nouvelles techniques de séquençage, le principe du pyroséquençage a été décrit pour la première fois en 1985. Il permet l'analyse d'une large variété de molécules, tels que l'ADN génomique, les produits PCR ou encore les ADN complémentaires. Il s'agit d'une méthode permettant d'analyser la synthèse d'ADN cible en temps réel. On parle ainsi de séquençage par synthèse d'ADN (Mardis, 2008). Avec le développement des techniques, le pyroséquençage permet le séquençage de fragments d'ADN de plus en plus longs. C'est notamment le cas de l'automate 454, développé par la société Roche, qui permettait au départ le séquençage de courts fragments d'ADN d'environ 100 puis 250 paires de bases (technologie GS-FLX) et qui aujourd'hui permet le

séquençage de fragments d'ADN de 350-400 paires de bases (Technologie GS-FLX-Titanium).

La méthode de séquençage Illumina est basée sur l'incorporation réversible de nucléotides fluorescents (CRT : *cyclic reversible termination*) et sur la lecture optique de la fluorescence. Comme pour la technique de Sanger, il s'agit d'une terminaison de synthèse (séquence supplémentaire attachée à la fin des lectures). Cette dernière est réalisée par l'utilisation d'un terminateur réversible contenant un groupement de protection attaché au dernier nucléotide incorporé lors de la synthèse d'ADN. L'élimination du groupement de protection, par photoclivage utilisant la lumière ultraviolette (> 300nm), permet la restauration du groupement fonctionnel du nucléotide incorporé, ce qui permet à l'ADN polymérase d'incorporer le prochain nucléotide et ainsi de suite. Il s'agit là également d'un séquençage en temps réel, basé sur la détection de la fluorescence, mais en présence des quatre nucléotides marqués (ce qui constitue un avantage par rapport à la technologie 454). En outre, le coût unitaire par paire de bases de cette technique est très largement moindre que celui du pyroséquençage. De plus, des études ont évalué quantitativement les erreurs dans les séquences consensus assemblées, et le pyroséquençage génère un taux d'erreur supérieur associé aux homopolymères de A et de T. Enfin, la couverture importante des séquences obtenue par Illumina grâce à son rendement élevé permet vraisemblablement de faciliter la résolution des ambiguïtés (Luo et al., 2012a). En revanche, les tailles de séquences unitaires sont plus petites que celles obtenues par 454 (36 puis 75 et maintenant jusqu'à 200 paires de bases).

Il est également à noter que ces évolutions ont généré une croissance considérable du volume des données acquises ce qui pose de plus en plus de problèmes de stockage et nécessite l'utilisation d'approches bioinformatiques complexes pour traiter cette information.

C'est ainsi que les organismes modèles puis, de plus en plus, non-modèles, ont vu leur génome séquencé, annoté et analysé au cours des années 1990 et 2000 (quelques exemples pour les eucaryotes, au génome de taille plus importante : *Saccharomyces cerevisiae* en 1996 (Goffeau et al., 1996), *Caenorhabditis elegans* en 1998 (C. elegans Sequencing Consortium., 1998), *Drosophila melanogaster* (Adams et al., 2000) et *Arabidopsis thaliana* en 2000 (The Arabidopsis Genome Initiative., 2000), *Homo sapiens* entre 2001 et 2006 (Gregory et al., 2006; McPherson et al., 2001; Venter et al., 2001), *Oryza sativa* en 2002 (Goff et al., 2002), *Populus trichocarpa* en 2006 (Tuskan et al., 2006). Cependant, ces données ne donnaient pas, dans un premier temps, accès au polymorphisme de séquence (un seul individu par espèce). Pour exploiter ces nouvelles approches dans des études s'intéressant à la diversité génétique et à l'évolution, il fallait étendre le séquençage à des échantillons au sein des populations.

2.2.1.2. Toujours plus de marqueurs : les SNP

L'accès au génome entier d'un individu étant de plus en plus facile, la comparaison de plusieurs

génomes d'une même espèce est rapidement devenue accessible. Dès lors, il est devenu possible d'accéder aux polymorphismes les plus fréquents au sein des génomes : les SNP, ou substitutions ponctuelle d'un nucléotide pour un autre. Leur détection repose généralement sur un principe simple : aligner tous les lectures générées par le séquençage d'individus différents sur un génome de référence (génome d'un individu donné, souvent choisi de manière arbitraire, assemblé et annoté) et en identifier les différences d'un nucléotide (en utilisant des algorithmes de tri pour éviter les faux positifs). On parle de reséquençage quand, après le séquençage d'un individu considéré comme le représentant de son espèce, de nouveaux individus sont séquencés pour obtenir le polymorphisme. La détection de ces SNP pour des études de génomique des populations, de santé humaine ou encore de recherche agronomique est de plus en plus répandu (Quillery et al., 2014; Uricaru et al., 2015; Xu et al., 2012).

Les SNP sont des polymorphismes généralement bi-alléliques, présents dans tout le génome (aussi bien les régions codantes que non-codantes) et très abondants puisqu'ils sont la forme principale de polymorphisme (90 % du polymorphisme humain est dû à des SNP [Erickson, 2003]). Leur répartition au sein des génomes n'est toutefois pas uniforme. L'intérêt pour des études de génomique des populations est d'avoir accès à de très nombreux marqueurs par individus permettant de localiser des régions génomiques précises (dans le cas où l'on veut identifier ces régions selon un critère donné) ou d'obtenir un panorama exhaustif de la variation génétique.

Cependant, la profondeur de séquençage lors du reséquençage d'un échantillon reste souvent moindre que celle lors de l'obtention de la séquence de référence (du fait de la couverture très importante requise pour obtenir un assemblage). Cette profondeur de reséquençage ne permet pas l'assemblage mais autorise la détection de la variation grâce à l'alignement à une séquence de référence. Ainsi, la détection des SNP reste dépendante de l'obtention d'un génome de référence sur lequel aligner les différentes lectures des séquençages. Chez certaines espèces non modèles, l'obtention de se génome de référence peut s'avérer difficile pour des problèmes de ploïdie, de séquences répétées ou de taille de génome. Des méthodes permettant de détecter les SNP sans génome de référence commencent à apparaître, et représentent une avancée pour les espèces non-modèles (Uricaru et al., 2015).

2.2.2. Les effets de la sélection sur le génome et les moyens de la détecter

La génomique des populations consiste à appliquer les principes de génétique des populations (voir section 2.1. et 2.2.1.) à des jeux de données génomiques (section 2.2.2.) afin d'identifier des régions du génome présentant des caractéristiques recherchées (les signes de sélection étant les plus courantes) ou de caractériser les patrons de variation génétique à l'échelle du génome entier. Cela nécessite de connaitre les effets attendus de ces forces sur les génomes pour pouvoir les détecter.

Dans cette partie nous nous intéressons aux statistiques mesurées en génomique des populations pour détecter la sélection.

2.2.2.1. La diversité nucléotidique et le polymorphisme

Une variable communément utilisée est la diversité nucléotidique, calculée comme le nombre moyen de différences nucléotidiques entre paires de séquence et notée π. C'est l'équivalent de l'hétérozygotie attendue à un locus *He*, et il s'agit d'un bon indicateur de la quantité de diversité existant au sein de la population. Dans le modèle neutre de l'évolution moléculaire le nombre de mutations entre deux séquences choisies au hasard dans la population vaut : T x μ x 2 avec T, le temps de coalescence entre ces deux séquences et μ, le taux de mutation. Le facteur de deux est dû au fait que les mutations apparaissent indépendamment sur les deux séquences. A chaque génération la probabilité d'un évènement de coalescence entre les deux séquences est $1/2Ne$ (la probabilité augmente si la population est petite, ce qui est une manifestation de la dérive plus forte). L'espérance d'un évènement de coalescence est ainsi de $2Ne$. L'espérance de π nommée θ vaut $4Ne\mu$ et correspond au taux de mutation de la population considérée de taille $2Ne$. La diversité nucléotidique est donc le produit du taux de mutation et de la taille efficace de la population. Autrement dit plus le taux de mutation est faible, ou plus la population est petite, et plus le polymorphisme attendu est faible.

Dans l'objectif de détecter de la sélection, une incompatibilité de π avec les valeurs attendues du modèle neutre pourrait paraitre très efficace mais il est très compliqué de comparer π à ses attendus neutres puisqu'il faudrait connaitre les paramètres *Ne* et μ. C'est dans ce contexte que d'autres statistiques ont été développées pour s'affranchir de ce problème.

Sous l'hypothèse nulle, la variation de polymorphisme entre locus ne dépend que de la variation du taux de mutation, variation qui peut être importante (Yang, 1996). Pour détecter la sélection à partir du polymorphisme il faut donc pouvoir discerner l'effet potentiel de la sélection de celui de la variation du taux de mutations. Le test HKA (du nom de ses auteurs Hudson, Kreitman et Aguadé) a été créé dans ce but précis (Hudson et al., 1987). Ce test part du principe que le taux de mutation neutre est constant dans le temps mais variable entre loci. Le test consiste en une utilisation d'une mesure de la divergence génétique, aux mêmes locus, entre deux populations (ou espèces). Le test statistique est assimilable à un test d'ajustement où les valeurs de diversité observées au sein et entre plusieurs populations ou espèces à plusieurs locus sont comparées aux attendus. Les attendus étant calculés sous cette hypothèse de proportionnalité entre polymorphisme intra et inter-populations.

2.2.2.2. La distribution des fréquences alléliques

Les tests basés sur la distribution des fréquences alléliques partent du constat que la sélection a pour effet de modifier les fréquences auxquelles ségrègent les allèles dans une population. Or la valeur de la diversité nucléotidique (π) est très sensible à la distribution des fréquences alléliques. Pour un nombre constant de sites polymorphes, plus les allèles des sites polymorphes sont en fréquences déséquilibrées (s'éloignant d'autant du 1:1), plus la diversité est faible. Or le nombre de sites polymorphes est, comme π, directement relié au taux de mutation populationnel θ. Deux estimateurs principaux de θ ont été créés, l'estimateur de Watterson (θs) et l'estimateur de Tajima ($\theta\pi$). θs est l'estimateur calculé à partir du nombre de sites polymorphes (Watterson, 1975). Sous le modèle de mutations en sites infinis, le nombre de sites polymorphes correspond au nombre de mutations. Les mutations se produisent donc sur des sites différents, hypothèse en général pas très éloignée de la réalité si l'on considère des SNP. $\theta\pi$ est l'estimateur calculé à partir du nombre moyen de différences entre deux séquences (Oleksyk et al., 2010; Tajima, 1983) soit la diversité nucléotidique décrite plus haut. Sous le modèle de Wright-Fisher, ces deux estimateurs sont égaux.

Cependant, si la population réelle viole une ou plusieurs des hypothèses (comme la neutralité sélective, l'association au hasard des allèles ou la taille constante), alors les deux estimateurs peuvent être différents. En effet, les écarts par rapport au modèle modifient drastiquement le spectre de fréquences alléliques. Par exemple, chaque variant rare constitue un site polymorphe mais contribue très peu à la diversité nucléotidique (π). Les deux estimateurs diffèrent donc par l'importance relative accordée dans le calcul aux variant rares et aux variants de fréquence intermédiaire. Le test de neutralité du *D* de Tajima est construit comme suit (Tajima, 1983) :

$$D = \frac{\theta\pi - \theta s}{\sqrt{\mathrm{var}(\theta\pi - \theta s)}}$$

Il permet de mesurer cet écart au modèle neutre de Wright-Fisher, tel que si $D<0$, c'est le signe d'un excès de variant rares et donc une preuve de sélection purificatrice ou balayage sélectif. Inversement, si $D>0$, c'est le signe d'un déficit de variants rares et donc une preuve de sélection balancée. Il est important de noter que les écarts au modèle neutre peuvent aussi être causés par la violation d'une autre hypothèse que la neutralité sélective (voir section 2.3).

Cette idée de construire un test comme le contraste entre deux estimateurs de θ fut reprise par la suite, avec des tests construits de manière à être sensibles à des types spécifiques de sélection naturelle. C'est le cas des tests *D* et *F* de Fu et Li (Fu and Li, 1993) ou *H* de Fay et Wu (Fay and Wu, 2000) qui peuvent présenter plus de sensibilité statistique que le *D* de Tajima dans le rejet de l'hypothèse de neutralité pour certains types de sélection (Depaulis et al., 2003). Plus généralement l'utilisation de la coalescence comme outil de simulation pour générer la distribution attendue des

statistiques se généralise en prenant en compte des modèles de plus en plus complexes (Li et al., 2003).

2.2.3.3. Mutations synonymes et non-synonymes

Les différentes méthodes présentées jusqu'ici portent majoritairement sur les effets de la sélection sur le polymorphisme au voisinage de la région sélectionnée mais ne permettent pas de quantifier les mutations réellement fixées par la sélection au sein des espèces. L'objectif des approches décrites dans cette section est de différencier, parmi les mutations fixées, lesquelles l'ont été par la sélection et par la dérive.

Ces tests se basent sur la redondance du code génétique qui veut que chaque mutation n'entraine pas forcément de changement de l'acide aminé sur lequel elle apparait. Lorsque qu'une mutation apparait sans changer l'acide aminé, on parle de substitution synonyme. Certaines positions des acides aminés sont entièrement synonymes, c'est le cas de la troisième base de la thréonine qui est indifférente (ACT, ACC, ACA et ACG codent pour la thréonine). Si, au contraire, l'acide aminé est modifié, on parle de substitution non-synonyme. La troisième base de la sérine est un bon exemple puisqu'elle est partiellement non-synonyme (AGT et AGC codent bien pour la sérine mais AGA et AGG codent pour l'arginine).

Un test courant concernant les mutations synonymes et non-synonymes est celui de McDonald et Kreitman (McDonald and Kreitman, 1991). L'abondance relative des polymorphismes non-synonymes et synonymes ségrégeant dans une population ($P(N)$ et $P(S)$ respectivement) mesure la proportion des variants non-synonymes dans les séquences codantes qui ne sont pas éliminés par la sélection purificatrice (en principe, $P(N)<P(S)$). Selon les hypothèses du modèle neutre, ce rapport est conservé si on considère le nombre de polymorphismes au sein de deux populations ou espèces ainsi que le nombre de substitutions entre elles. En effet, toujours sous le modèle neutre, le rapport de mutations synonymes et non-synonymes dans la divergence des deux populations (ou espèces) $d(N)$ / $d(S)$ est lui aussi égal à la proportion de mutations qui sont neutres et qui sont pas éliminés par la sélection purificatrice. Le test de McDonald et Kreitman revient donc à comparer le ratio de substitutions synonymes et non-synonymes au sein d'une espèce en comparaison à une autre espèce, ou population (au même locus). Les écarts par rapport à cette hypothèse montrent un épisode de sélection positive ($d(N)/d(S)>P(N)/P(S)$) ou le maintien de mutations délétères au sein des populations (théorie dit "quasi neutre" (Ohta, 1973; Ohta and Gillespie, 1996)) (PN/PS>DN/DS) à cause des variants faiblement délétères qui se maintiennent transitoirement au sein des espèce) (Stukenbrock and Bataillon, 2012).

2.3. Les effets confondants

2.3.1. Les effets démographiques

La démographie est l'ensemble des propriétés des populations (taille actuelle et passée, incluant les changements de taille ancestraux, structuration et patrons de migration). Les phénomènes démographiques les plus importants sont les changements de taille de populations, la sous-structuration des populations, les phénomènes de mélanges entre populations. Parmi les changements de taille de population, le goulot d'étranglement correspond à une réduction de la taille population suivi d'une expansion. Celui-ci peut être dû à une modification environnementale mais aussi résulter d'un évènement sélectif de grande ampleur (saut d'hôte ou contournement de résistance dans le cas de pathogènes). La force et l'âge des goulots d'étranglement sont des paramètres essentiels pour ce type d'évènement. A l'inverse, l'expansion démographique correspond à un accroissement de la taille de la population, ce qui peut résulter de la colonisation d'un nouvel environnement (effet de fondation).

L'interprétation des déviations des tests de neutralité se base sur l'hypothèse de stabilité démographique. Si la démographie des populations n'est pas prise en compte lors de l'application du test de neutralité, de nombreux faux positifs/négatifs seront imputables à la démographie. En effet les phénomènes démographiques décrits ci-dessus sont capables de modifier les probabilités de coalescence au cours du temps et donc modifient la généalogie des séquences par rapport à celle attendue sous le modèle standard. Or ce sont ces modifications qui sont mises en évidence par le *D* de Tajima, par exemple. Ainsi un goulot d'étranglement sévère ou une expansion démographique créent des arbres de coalescence aux branches internes très courtes par rapport aux branches terminales et donc des *D* de Tajima négatifs qui pourraient être interprétés comme des évènements sélectifs. Au contraire, les phénomènes réduisant temporairement la taille de population (tels que les goulots d'étranglement récents et de force modérée) engendrent des *D* positifs (Depaulis et al., 2003).

Il est donc indispensable de connaitre l'histoire démographique des populations étudiées pour générer une distribution empirique des statistiques adaptée aux modèles étudiés. Cette distribution permettra des tests plus robustes des effets de la sélection car on réduit fortement les effets de la démographie (Siol et al., 2010).

2.3.2. Les modes de reproduction

Ce travail de thèse est centré sur l'étude d'un champignon. Les champignons présentent une très large diversité de modes de reproduction (Jin et al., 2010; Taylor et al., 1999). Certaines espèces se reproduisent de manière purement sexuée, d'autres de manière purement asexuée et d'autres, enfin, alternent entre les deux modes. Mais, même au sein d'une espèce, certaines populations peuvent

présenter des modes de reproduction différents (Goyeau et al., 2007; Xhaard et al., 2011). Or, le mode de reproduction peut jouer un rôle majeur dans la structure de la diversité et les caractéristiques du polymorphisme génétique d'une espèce. De ce fait, un grand nombre de statistiques mesurables en génétique des populations sont impactées par le mode de reproduction de l'espèce.

Pour les espèces réalisant de la reproduction sexuée, la recombinaison brise le déséquilibre de liaison entre les différents sites du génome. De ce fait, l'histoire évolutive de séquences séparées par une distance génétique suffisante deviennent indépendantes. L'effet de la sélection se limite donc aux régions génétiquement liées à la région sélectionnée. Les tests de neutralité deviennent dès lors plus conservatifs, rejetant moins souvent l'hypothèse de neutralité à distance de la région sélectionnée, du fait de la portée limitée de la sélection (Wall, 1999).

Le deuxième effet de la recombinaison concerne la sélection positive. La sélection positive tend à éliminer du polymorphisme au niveau populationnel et seules les mutations récentes (postérieures à l'évènement de sélection) sont visibles. Mais ce cas de figure n'est vrai que si c'est l'ancêtre commun de toute la population considérée qui a subi cette mutation sélectionnée. Or, en cas de recombinaison, la fixation de l'allèle favorable peut se limiter à une fraction de la séquence (ce qui est lié au locus sélectionné) et donc n'induire qu'une élimination partielle du polymorphisme (Kaplan et al., 1989). Cet effet crée dans la généalogie une coexistence de coalescences récentes (celle du locus sélectionné et des locus qui lui sont liés) et anciennes (les locus ayant échappé à la sélection par recombinaison). Ce type de phénomène induit des signatures spécifiques dans le polymorphisme (Fay and Wu, 2000).

Les espèces se reproduisant de manière asexuée vont présenter d'autres patrons susceptibles de biaiser les tests. Tout d'abord, le déséquilibre de liaison entre séquences est plus fort. En cas de sélection, l'effet peut se propager sur le chromosome entier. Ce phénomène, appelé auto-stop génétique, rend difficile l'identification de locus particuliers puisque que les patrons de diversité sont homogénéisés. En plus de cela les populations asexuées présentent classiquement des patrons de diversité particuliers, avec par exemple un taux d'hétérozygotie très élevé (et donc un F_{IS} très négatif) ainsi qu'un grand nombre de génotypes répétés (Balloux et al., 2003; Halkett et al., 2005). L'ensemble de ces phénomènes font que les populations asexuées sont difficilement étudiables avec les modèles classiques basés sur le polymorphisme (prévus pour des populations sexuées).

Il est important de prendre en compte le mode de reproduction des populations étudiées pour éviter les biais cités ci-dessus. Heureusement, les indices de génétique des populations permettent de détecter le mode de reproduction d'une population (Bourassa et al., 2007; Jin et al., 2010; Xhaard et al., 2011). Mais peu de méthodes existent pour prendre en compte ces modes de reproduction alternatifs dans la recherche de sélection alors qu'ils peuvent engendrer des faux positifs (Siol et al.,

2010).

2.3.3. NGS, de nombreux faux positifs

Toutes les méthodes visant à étudier la démographie ou à détecter la sélection reposent sur la détection du polymorphisme au sein des séquences, partant du principe que ce polymorphisme est réel. Or les techniques de séquençage de nouvelle génération (NGS) présentent un taux d'erreur non négligeable, supérieur au séquençage Sanger, dont le type diffère selon les techniques utilisées (tableau 1) (Knief, 2014). Ces erreurs peuvent créer des faux positifs, des SNP sans existence réelle. Le taux est généralement compris entre 0,4 et 1 % d'erreur, toute technologie confondue. Partant d'un taux faible de 0,4 % d'erreur, en considérant un génome de 100 Mb environ (ce qui sera le cas du modèle d'étude de ce travail) séquencé en 20X (donc 2000 Mb de données), cela représente 8 000 000 bases erronées. Des études se sont intéressées aux différentes sources d'erreur selon la technologie :

Avec un taux d'erreur de 0,5 %, chaque lecture de 200 pb devrait contenir une erreur en moyenne mais la répartition de ces erreurs n'est pas homogène le long des génomes. Pour la technologie 454, 70 % des lectures sont considérées comme sans erreur et entre 57 % et 76 % pour Illumina (selon la plateforme) (Huse et al., 2007; Quail et al., 2012). Une des explications envisagées est la présence de librairies multiples lorsque les séquences sont semblables, ce qui est souvent le cas des éléments transposables (Huse et al., 2007).

Le deuxième type d'erreur le plus fréquemment rencontrée correspond aux insertions et délétions (indel). Mais encore une fois leur fréquence dépend de la technologie utilisée. Les insertions de bases représentent l'erreur la plus fréquente pour la technologie 454, suivie de la délétion d'une base. La majorité de ces erreurs se situent dans des régions homopolymériques et s'expliquent par la précision de la détection du signal lumineux qui diminue avec l'augmentation du nombre de bases identiques (Margulies et al., 2005). A contrario, les substitutions (remplacement d'une base par une autre) représentent le type d'erreur le plus fréquemment rencontré pour la technologie Illumina (Nguyen et al., 2011). Le plus faible taux d'erreur de type indels de la technologie Illumina par rapport à la technologie 454 est dû à la stratégie de blocage terminal au cours du processus de séquençage, ce qui permet l'incorporation d'une seule base par cycle de séquençage, de sorte qu'une région homopolymérique est séquencée base par base.

Les erreurs de séquençage ont tendance à s'accumuler à la fin des lectures, ce qui est associée à une réduction de la qualité des bases. Cette accumulation d'erreurs est le résultat d'une diminution du rapport signal sur bruit (le bruit de fond devient équivalent au signal recherché) au cours du processus de séquençage, qui détermine en grande partie la longueur maximale de lecture de toutes les plateformes de séquençage. Cette accumulation devient problématique à partir de 200

Compagnie	Taux d'erreur [%]	Données brutes	Référence
Roche 454	1	Non	Glenn, 2011
	1,1	oui	Gilles et al., 2011
	4	oui	Margulies et al., 2005
	< 1	Non	Thompson and Milos, 2011
	0,25	oui	Huse et al., 2007
	0,4	oui	Quinlan et al., 2008
	1,1	oui	Lind et al., 2010
	0,4 – 0,5	oui	Niu et al., 2010
	0,4	oui	Quince et al., 2011
	0,11 – 0,34	oui	Vandenbroucke et al., 2011
	~0,4	oui	Loman et al., 2012
	0,46	oui	Jünemann et al., 2013
Illumina	0,5	Non	Mardis, 2013
	<0,1	Non	Glenn, 2011
	<2	Non	Liu et al., 2012
	1 – 1,5	Non	Shendure and Ji, 2008
	< 1	Non	Thompson and Milos, 2011
	0,6 - 1	oui	Dohm et al., 2008
	1,3	oui	Hillier et al., 2008
	0,26 – 0,80	oui	Quail et al., 2012
	< 0,8	oui	Quail et al., 2008
	2,5 – 7,3	oui	Minoche et al., 2011
	5,2 – 6,0	oui	Nguyen et al., 2011
	~ 0,1	oui	Loman et al., 2012
	0,09	oui	Jünemann et al., 2013

Tableau 1 : Taux d'erreur de séquençage dans les articles originaux (données brutes) et les revues en fonction de la technologie utilisée. Adapté de Knief, (2014).

à 300 pb pour la technologie 454 et à partir de 100 à 150 pb pour la technologie illumina, expliquant les différences de capacités en terme de longueur de séquence entre ces deux technologies (Gilles et al., 2011; Kircher et al., 2009)

Le pourcentage en GC des lectures semblent également introduire un biais dans le séquençage. Les analyses montrent une sous-représentation des régions riches en AT tout comme les régions riche en GC, au bénéfice des régions intermédiaires (Bentley et al., 2008; Kozarewa et al., 2009; Quail et al., 2012).

Une fois les types d'erreur connus, un certain nombre de méthodes ont été développées dans le but de les corriger, ou du moins de les compenser. La première d'entre elles consiste à augmenter la profondeur de séquençage, permettant de compenser les erreurs par la quantité de lectures de la même région (Margulies et al., 2005). Une deuxième stratégie consiste à utiliser plusieurs plateformes générant des erreurs différentes, identifiables par comparaison et donc éliminables (Koren et al., 2012; Nakamura et al., 2011). Pour finir, une stratégie efficace dans la réduction des erreurs de lecture consiste à prendre en main l'étape d'identification des bases (base calling , algorithme qui décide de la base lue en fonction des données optiques obtenues par l'automate de séquençage), avec des algorithmes présentant des performance supérieures aux algorithmes standards (Das and Vikalo, 2013). Pour finir, une dernière consiste à prendre en main l'étape de SNP calling (algorithme retenant ou pas les polymorphismes selon des critères de qualité). En effet l'utilisation de paramètres différents lors de cette étape peut faire varier énormément le nombre de SNP retenus et donc le niveau de polymorphisme (voir exemple de *Puccinia striiformis* f. sp. *tritici* dans Duplessis et al., (2014).

3. Modèle d'étude : *Melampsora larici-populina*, agent de la rouille du peuplier

Melampsora larici-populina est un des agents majeurs de la rouille du peuplier. Cette espèce appartient à la division des Basidiomycètes, à la classe des Pucciniomycètes et à l'ordre des Pucciniales. L'ensemble des rouilles appartient à l'ordre des Pucciniales et sont un groupe de pathogènes des plantes capables d'infecter une grande diversité d'hôte. Elles sont responsables de quelques-unes des pertes économiques les plus importantes pour l'agriculture telles que celles affectant le blé, le café, le soja, les pins, le peuplier et l'eucalyptus.

Le genre *Melampsora* compte plus d'une cinquantaine d'espèces dont 17 sont pathogènes sur les peupliers (genre *Populus*) (Vialle et al., 2011). Sur les neuf espèces de *Melampsora* présentes en Europe, *M. larici-populina* est celle qui cause les dégâts les plus importants dans les peupleraies françaises (Pinon and Frey, 2005). L'infection des peupliers par la rouille a des

conséquences considérables sur ces derniers mais c'est moins le cas pour son hôte alternatif, le mélèze *Larix decidua*.

3.1. Un cycle de vie complexe

Melampsora larici-populina présente un cycle complexe puisque c'est une rouille hétéroïque (qui accomplit son cycle sur deux plantes hôtes d'espèces différentes) et macrocyclique (cinq types de spores sont produits lors du cycle). La reproduction sexuée a lieu sur l'hôte écidien (le mélèze) au printemps. Elle est suivie de plusieurs cycles de multiplication asexuée sur l'hôte télien (le peuplier) du printemps jusqu'à l'automne (figure 6).

A l'automne, lors de la sénescence des feuilles de peuplier, se forment les télies contenant les téliospores dicaryotiques. C'est la forme de résistance sous laquelle *M. larici-populina* passe l'hiver. La caryogamie (fusion des noyaux, marquant le début de la phase sexuée) a lieu au sein des téliospores, générant les basides diploïdes qui subissent au printemps, une méiose produisant ainsi des basidiospores haploïdes qui infectent les aiguilles de mélèze. Ces primo-infections génèrent des spermogonies, qui produisent des spermaties. Une plasmogamie entre deux spermaties de signes sexuels opposés engendre des écidies dicaryotiques qui produisent des écidiospores disséminées par le vent jusqu'aux feuilles de peuplier. Cette infection génère des urédies sur la face inférieure des feuilles. Ces urédies se propagent alors par multiplication asexuée et peuvent infecter d'autres feuilles par production d'urédiniospores et ce jusqu'à la fin de l'automne.

Chez *M. larici-populina*, le processus d'infection a particulièrement bien été étudié au cours de la phase urédienne (Hacquard, 2010; Hacquard et al., 2012; Laurans and Pilate, 1999; Rinaldi et al., 2007). Une fois en contact avec la feuille, les urédiniospores produisent un tube germinatif qui, une fois en contact avec un stomate, arrête sa croissance et se différencie en une structure appelée appressorium, à partir duquel sera émis un hyphe primaire d'infection permettant la pénétration de l'agent pathogène au sein des tissus foliaires par le stomate. En quelques heures, le champignon développe une vésicule sous-stomatique à partir de laquelle se différencie un hyphe qui se propage entres les cellules du mésophile. Au contact de celles-ci, l'extrémité de l'hyphe différencie une cellule-mère haustoriale. Celle-ci permet la formation d'une structure d'infection au niveau de la cellule-hôte, l'haustorium, qui est une invagination dans la membrane plasmalemmique au-delà de la paroi cellulaire, et autour de laquelle une matrice spécifique est formée. Cette structure, permet le détournement des métabolites de la plante et la sécrétion de molécules effectrices dans l'apoplasme et dans le cytoplasme de la cellule hôte (Catanzariti et al., 2007). La différenciation croissante des haustoria dans les premiers jours après l'infection va assurer une croissance rapide au champignon. Après quelques jours (environ une centaine d'heures), le mycélium forme des cellules sporifères au sein de l'espace intercellulaire situé sous l'épiderme inférieur de la feuille à partir desquelles vont

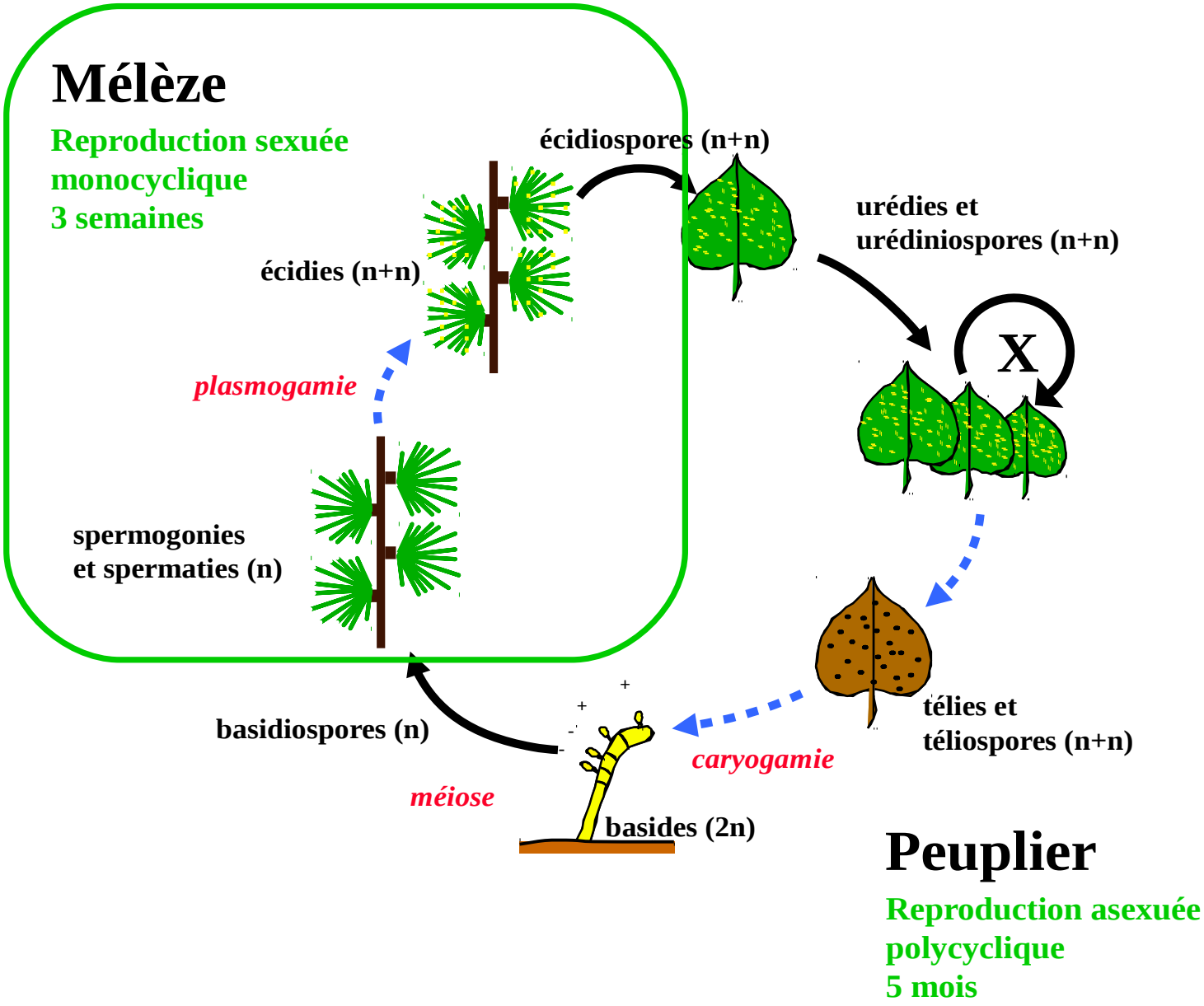


Figure 6 : Cycle biologique de *Melampsora larici-populina* comprenant ses deux hôtes, le peuplier (hôte télien) et le mélèze (hôte écidien).
Les flèches bleues et noires indiquent respectivement des changements d'état et des évènements de dispersion. D'après Pinon and Frey (2005).

se différencier les urédies visibles à la surface des feuilles (amas de spores orangées) au bout d'une semaine environ.

3.2. Un problème écologique et économique

Sur les peupliers, *M. larici-populina* affecte les capacités photosynthétiques de l'arbre et ainsi engendre une mauvaise lignification des rameaux, une faible croissance de l'arbre, un retard dans le développement racinaire et enfin une décoloration et une nécrose des feuilles, suivie d'une éventuelle défoliation précoce (Gerard et al., 2006). Le peuplier est la troisième essence de feuillus en termes de récolte en France grâce à sa croissance rapide. Les peupleraies sont surtout présentes dans le Nord de la France. La sélection variétale a privilégié jusqu'à présent la résistance qualitative à la résistance quantitative. Or une fois la résistance qualitative contournée par le champignon, la plupart des peupliers cultivés exprimant une très faible résistance quantitative, les symptômes sont particulièrement intenses. De plus, les peupleraies sont majoritairement composées d'un seul cultivar de peuplier (il s'agit même de clones en raison des techniques de multiplication par bouturage). *Melampsora larici-populina* ne fait pratiquement pas de dégâts chez les peupliers sauvages du fait de leur résistance quantitative plus élevée et de leur diversité génétique qui empêche la susceptibilité simultanée de grandes populations. En revanche, depuis les années 1990, et suite à la diffusion intensive de certains cultivars (clones) de peuplier (dont par exemple Unal et surtout Beaupré), auxquels il s'est adapté dès 1994, *M. larici-populina* fait maintenant d'importants dégâts sur peupleraies cultivées (allant de fortes défoliation avec retards de croissance à la mort des arbres).

A ce jour, huit contournements de résistances distinctes ont été décrits chez *M. larici-populina* (de R1 à R8). Les évènements les plus marquants ont été les contournements des résistances qualitatives R2 (portée principalement par le cultivar Luisa-Avanzo) en 1986, R7 (portée majoritairement par Beaupré) en 1994 et R8 (portée par Hoogvorst) en 1997.

Afin de caractériser les virulences des isolats de *M. larici-populina* collectées sur le terrain, une gamme différentielle de cultivars, présentant des résistances contrastées vis-à-vis de la maladie a été établie à l'INRA de Nancy dans l'équipe d'Ecologie des Champignons Pathogènes Forestiers par le groupe de Pascal Frey (Gerard et al., 2006; Pinon and Frey, 1997). Les isolats peuvent ainsi être caractérisés par leurs aptitudes à infecter un ou plusieurs cultivars de cette gamme et la combinaison de virulences qui caractérise un isolat est appelée pathotype. En théorie, avec huit virulences, on peut définir 256 pathotypes différents. Néanmoins à ce jour, seuls 57 pathotypes ont été identifiés dans la nature (Pinon and Frey, 2005).

Le contournement du gène de résistance R7 par *M. larici-populina* a entrainé des épidémies de rouille sans précédent et causé des pertes conséquentes pour la populiculture française depuis

1994 (Pinon and Frey, 2005). Même si la date exacte du contournement n'est pas connue, les premières traces d'individus virulents 7 sont observées dès 1994 et, grâce à un suivi épidémiologique, une augmentation drastique de l'aire de répartition et de la prévalence de la maladie dans les plantations de peupliers a été observée quelques années après. Ainsi tout se passe comme si le ou les gènes conférant la virulence 7 de *M. larici-populina* s'étaient propagés en France et, ainsi, on retrouve cette virulence sur quasiment tout le territoire depuis 1999 (figure 7).

C'est sans doute grâce à la pression de sélection engendrée par la plantation massive du même cultivar sur le territoire (en 1996 les clones de peupliers porteurs de la résistance 7 représentaient 60% des ventes avec 1 200 000 unités contre moins de 200 000 en 2006 ; Hayden et al ; non publié) que cette explosion démographique a pu avoir lieu. Cette sélection forte, à une échelle de temps courte (1994-2011), des individus contournants a sans doute laissé des traces génétiques et génomiques. A la suite d'un contournement de gène de résistance, une différenciation entre populations virulentes et avirulentes peut apparaître chez l'agent pathogène (Guerin et al., 2007). Les individus virulents sont souvent moins diversifiés génétiquement que les individus avirulents. Ceci est du à l'effet de fondation qui suit le goulot d'étranglement se produisant au moment du contournement et conduisant à une absence de recombinaison. En effet, la population virulente subit les effets combinés de la dérive (petit nombre d'individu initial) et de la pression de sélection exercée par l'hôte. Ainsi l'étude de populations contournantes permet de suivre à la fois les conséquences démographiques de l'adaptation d'agents pathogènes à leurs hôtes. Tout ceci fait de l'étude de ce contournement un bon exemple de l'évolution des interactions plante-pathogène et des conséquences génétiques et génomiques d'évènements drastiques de sélection.

3.3. Structure des populations de *M. larici-populina*

La structuration et l'évolution des populations de *M. larici-populina* sont fortement impactées par la sélection exercée par les cultivars de peupliers présents sur le territoire (Gerard et al., 2006; Pinon and Frey, 1997; Xhaard et al., 2011). Le développement d'un ensemble de 25 marqueurs microsatellites spécifiques à *M. larici-populina* a permis d'envisager des études de génétique des populations (Xhaard et al., 2011). Une étude portant sur les profils de virulence et le génotypage de 476 individus, échantillonnés sur tout le territoire français en 2009, a montré une différence nette entre les régions du Nord et du Sud de la Loire. Les individus portant la virulence 7 (Vir7) s'observant en grande majorité au Nord et les individus avirulents (Avr7) généralement au Sud de cette limite (Xhaard et al., 2011). Les plantations du cultivar 'Beaupré', porteur de la résistance qualitative R7, ont été principalement réalisées au Nord de la Loire et coïncident donc avec la répartition du groupe génétique des Vir7. Cela indique une influence des cultivars de peupliers (sans doute par leurs gènes de résistances) sur la structuration et l'évolution des populations. Un autre

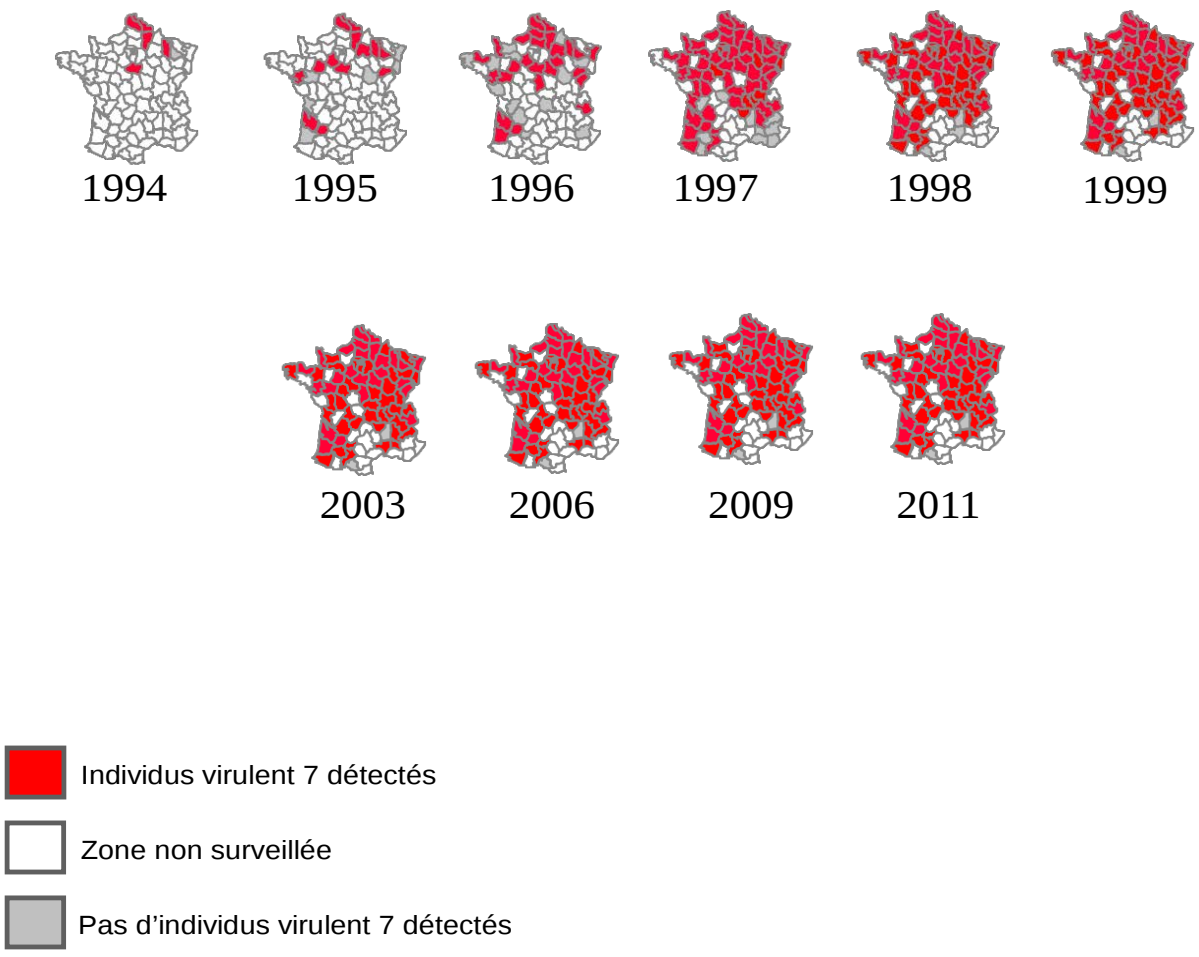


Figure 7 : Evolution de l'incidence des isolats de *Melampsora larici-populina* virulent 7 en France de 1994 à 2011. D'après Xhaard et al., (2011).

groupe génétique a été identifié au cours de cette étude. Il s'agit d'individus retrouvés essentiellement dans le Sud de la France et présentant toutes les caractéristiques d'une reproduction purement asexuée : richesse allélique faible, excès d'hétérozygote et F_{IS} très négatifs (Balloux et al., 2003; Halkett et al., 2005). Ce groupe génétique ne fait donc peu ou pas de reproduction sexuée sur mélèze et est ainsi complètement inféodé aux peupliers.

Des travaux de Hayden et al. (non publiés), ont par la suite confirmé l'influence de la répartition des cultivars de peuplier (et donc des résistances qu'ils portent) sur les profils de virulence observés chez *M. larici-populina*. Une comparaison entre les densités de plantation des différentes résistances qualitatives par région à l'échelle de la France et les profils de virulence de populations de *M. larici-populina* a été réalisée. Les résultats montrent une corrélation forte entre le taux de plantation du cultivar 'Beaupré' (porteur de la résistance R7) et la fréquence de la virulence 7 à l'échelle du territoire. Le résultat est identique pour le couple résistance/virulence 8. Ces résultats illustrent l'évolution des groupes génétiques de *M. larici-populina* en réponse aux pressions de sélection exercées par les gènes de résistance déployées. Cependant, aucune corrélation n'a été mise à jour pour les couples résistance-virulence 1, 3, 4 et 5. Les fréquences de ces quatre virulences sont corrélées à la fréquence de la virulence 7, qui a probablement émergée dans un fond génétique où ces virulences étaient présentes.

3.4. Caractéristiques des génomes de rouilles

Les premiers génomes d'espèce de rouilles a avoir été séquencés ont été ceux de *M. larici-populina* et de l'agent de la rouille noire du blé *Puccinia graminis* f. sp. *tritici* (Duplessis et al., 2011a; McDowell, 2011). Le séquençage du génome de *M. larici-populina* (souche 98AG31, un isolat virulent 7 devenu dès lors souche de référence) a été initié dès 2006 par l'UMR 1136 du Centre INRA de Nancy et un consortium international en collaboration avec le centre international de séquençage Joint Genome Institute du département de l'énergie américain. Ces deux espèces sont des rouilles au cycle de vie similaire (rouilles macrocycliques hétéroïques). Ces deux rouilles présentent des organisations génomiques comparables avec un génome de grande taille (80 et 101 Mb), une grande concentration en éléments transposables (environ 50 %) et un grand répertoire de gènes (supérieur à 16 000 gènes). Les autres champignons biotrophes obligatoires pathogènes de plantes qui ont été séquencés ne partagent qu'un nombre restreint de ces caractéristiques (Duplessis et al., 2013). Ainsi, ces caractéristiques pourraient être corrélées au statut de rouille hétéroïque macrocyclique de ces deux espèces séquencées et ne pas être conservées entre toutes les rouilles. Le génome de la rouille du lin *Melampsora lini*, qui est une rouille autoïque, fait 189 Mb et porte 16 271 gènes putatifs avec 45 % de TE (Nemri et al., 2014). De plus la rouille microcyclique *Phakopsora pachyrhizi*, responsable de la rouille du soja, possède un génome de taille estimée entre

500 Mb et 1 Gb selon le degré d'hétérozygotie avec 19 000 gènes prédits (Loehrer et al., 2014). Ces données suggèrent que la grande taille du génome des rouilles est un trait commun vraisemblablement dû à l'expansion de famille d'éléments transposables mais ne serait pas lié aux différences de cycle de vie (Duplessis et al., 2014).

La grande taille et la richesse en TE des génomes des rouilles rend leur séquençage et leur assemblage très complexe. Le séquençage du génome de l'agent de la rouille jaune du blé *Puccinia striiformis* f. sp. *tritici* en est une bonne illustration. Une première ébauche de génome de référence de 65 Mb a été réalisée avec un séquençage Illumina de grande profondeur mais de faible résolution (Cantu et al., 2011). Cet assemblage a permis la prédiction d'un catalogue de 20 423 gènes putatifs, plus conséquent que celui de l'espèce proche *P. graminis* f. sp. *tritici* (Cantu et al., 2011). Un autre projet de séquençage et d'assemblage du génome a été réalisé sur un isolat chinois de *P. striifformis* f. sp. *tritici* à l'aide d'une méthode « fosmide à fosmide » combinant les technologies Illumina et Sanger. Cette méthode a permis l'obtention d'un meilleur assemblage associé à une profondeur de séquençage supérieure. La taille du génome obtenu est de 110 Mb avec 25 288 gènes prédits (Zheng et al., 2013). La stratégie employée par Zheng et al. pourrait être prometteuse pour l'étude du génome d'espèces non modèles, riches en TE (par exemple *Blumeria* spp, (Hacquard et al., 2013; Spanu et al., 2010; Wicker et al., 2013)). L'obtention d'une ébauche de génome à moindre coût avec une approche plus directe, telle que celle utilisée par Cantu et al.(2011) reste très intéressante pour un accès rapide au polymorphisme de pathogènes importants sur le plan agronomique (Duplessis et al., 2014).

Des analyses comparatives des génomes des agents des rouilles ont permis de mettre à jour des familles de gènes spécifiques des Pucciniales plus abondantes que chez d'autres basidiomycètes comme les hélicases, les protéines riches en cyctéines, les transporteurs d'oligopeptides et les lipases ou les peptidases (Duplessis et al., 2011a; Zheng et al., 2013). Cette abondance spécifique peut s'expliquer par la relation à l'hôte ou dans le cas des hélicases et d'autres gènes associés à la maintenance ou la réparation de l'ADN une signature spécifique du génome des rouilles suite à l'envahissement par des TE au cours de son histoire évolutive. Les catalogues de gènes ainsi que l'organisation des génomes sont relativement différentes entre les Pucciniaceae et les Melampsoraceae, ce qui peut sans doute s'expliquer par la divergence ancienne de ces deux familles (Aime et al., 2006; McTaggart et al., 2015).

L'accès aux méthodes de séquençage des génomes étant de plus en plus aisé, les recherches de polymorphisme au sein de ces espèces est devenu un axe de recherche important. Ainsi chez *P. striiformis* f. sp. *tritici* deux études ont permis d'identifier respectivement 109 000 et 350 000 SNP en moyenne par individu indiquant un niveau de diversité élevé chez cette espèce (Cantu et al., 2013; Zheng et al., 2013). Les différences entre les deux études peuvent s'expliquer par l'utilisation

d'individus différents, la qualité initiale des génomes de références considérés, les algorithmes de tri et les paramètres utilisés pour détecter les SNP. En comparaison, 88 083 et 129 172 SNP ont été rapportés au sein des génome de référence chez *M. larici-populina* et *P. graminis* f. sp. *tritici* respectivement (Duplessis et al., 2011a).

3.5. Annotation fonctionnelle et transcriptomique

La transcriptomique est une technique qui a pour objet l'étude de l'ensemble des transcrits d'un organisme à différents niveaux (organes, tissus, cellules). Les techniques les plus communément utilisées à l'heure actuelle pour réaliser de telles analyses sont les méthodes de séquençage de nouvelle génération présentées précédemment et appliquées à un ensemble de transcrits. Il y a encore quelques années, notamment lors des premiers projets de génomique sur les agents des rouilles, les puces à oligonucléotides (de type Affymetrix ou NimbleGen par exemple) étaient plus particulièrement utilisées.

L'annotation automatique du génome de *M. larici-populina* par le JGI et l'annotation experte par le consortium international qui a réalisé l'analyse du génome a révélé la présence 16 399 prédictions de gènes dans un génome de 101 Mb (Duplessis et al., 2011a). Parmi ces gènes, 5798 étaient spécifiques de *M. larici-populina* et répartis en 909 familles de gènes (Hacquard et al., 2011). Ce niveau de spécificité génétique est toutefois à nuancer par le nombre faible de génomes de rouilles disponibles en 2011 dans les bases de données. Notamment le séquençage récent de l'agent de la rouille du lin *M. lini* a permis de montrer qu'une large proportion de ces gènes sont conservés au sein des Melampsoraceae (Nemri et al., 2014).

Les travaux de thèse de Stéphane Hacquard conduits au sein de l'UMR 1136 Interactions Arbres/Micro-organismes ont permis de réaliser une analyse globale du sécrétome ainsi que du transcriptome de *M. larici-populina* 98AG31. Un total de 1184 gènes (représentant 7,2 % des gènes) codant des petites protéines sécrétées (SSP) ont été décrits au moyen de l'annotation experte du génome et d'analyse de l'expression de ces gènes. Au total, 169 familles de gènes codant des SSP ont été décrites (la plus grande famille contenant 111 membres). Ces SSP, qui présentent des caractéristiques typiques des effecteurs décrits jusqu'à présent chez les champignons, pourraient potentiellement être impliquées dans la pathogénicité de *M. larici-populina* (Duplessis et al., 2011a; Hacquard et al., 2012). Ces travaux ont permis d'établir une liste de ces SSP, la position des gènes correspondants et leurs organisations en cluster ou en tandem au sein du génome, leur richesse en cystéines, leurs homologies dans les bases de données, ainsi que leurs classements en familles géniques.

Une analyse du transcriptome de *M. larici-populina* a été réalisée pour suivre l'expression de ces gènes au cours de l'infection de feuilles du cultivar Beaupré de peuplier (à 2, 6, 12, 24, 48,

96, 168 heures post-inoculation), ainsi que dans les urédiniospores dormantes et les urédiniospores germées (Duplessis et al., 2011b). Cette analyse a identifié des classes de SSP en fonction de leurs profiles d'expression au cours du cycle infectieux, avec parmi elles des gènes présentant une expression précoce (de la formation des premiers haustoria jusqu'à 24-48 heures après inoculation), intermédiaires (lors de la phase de croissance biotrophe entre 48-96 heures) et tardives (croissance biotrophe et formation des nouvelles urédiniospores entre 96-168 heures). Les classes de gènes fortement exprimés en début d'infection représentent des candidats sérieux pour la recherche des déterminants de la manipulation des systèmes de défenses de l'hôte, en vue de favoriser la virulence (Duplessis et al., 2011b).

4. Stratégie de l'étude

L'objectif central de ce travail est d'identifier l'impact des contournements de résistance sur le champignon phytopathogène *M. larici-populina*, de l'histoire démographique au déterminisme génétique. Dans ce but, des approches de génomique comparative, de génétique/génomique des populations et de biologie évolutive ont été employées.

Dans le chapitre 2, une étude préliminaire basée sur le séquençage de 15 isolats de *M. larici-populina* nous a permis d'estimer l'impact des contournements de résistance sur le génome de *M. larici-populina*. Pour cela, les 15 isolats ont été sélectionnés pour leurs profils de virulence distinct et leur génome a été séquencé. L'objectif était de savoir s'il était possible d'identifier de la sélection à cette échelle intra-spécifique, et si nous pouvions corréler cette sélection aux profils de virulence dans le but d'identifier des effecteurs candidats.

Dans un troisième chapitre, nous avons étudié l'impact démographique d'un évènement majeur, le contournement de la résistance 7, sur la structure génétique des populations de *M. larici-populina* en France. Une étude antérieure a permis de montrer que cet évènement avait largement façonné la structure génétique des populations de ce pathogène, avec une forte différenciation entre des individus avirulents 7 (incapables de contourner la résistance 7 et originaires de la zone de distribution originelle du Peuplier noir) et des individus virulents 7 (Xhaard et al., 2011). Au sein de la collection historique disponible au laboratoire, 600 isolats de *M. larici-populina* ont été choisis avec une maximisation de leur répartition historique (de 1992 à 2012) et géographique (Nord, Est et Sud de la France). Ils ont été génotypés à l'aide de 21 marqueurs microsatellites. Nous avons conduit une étude de génétique des populations dans le but de répondre aux questions suivantes : Combien de groupes génétiques expliquent la diversité génétique au sein des et entre les isolats ? Quelles sont les caractéristiques de ces groupes génétiques ? Combien de groupes ont été impactés par le contournement de la résistance 7 ?

Le quatrième chapitre cherche à identifier les balayages sélectifs, et les gènes ou les régions génomiques soumises à sélection lors de ce contournement, en tenant compte des effets confondants des fluctuations démographiques. Pour remplir ces objectifs nous avons sélectionné quatre populations clés tenant compte des résultats sur la structuration populationnelle du deuxième chapitre. Au total, 86 génomes ont été séquencés au sein de ces quatre populations et le polymorphisme a été détecté par alignement sur le génome de référence. Une approche d'inférence démographique basée sur des simulations de coalescence a permis de déterminer le scénario démographique le plus probable, et d'en tirer les distributions attendues des statistiques. Pour finir, un scan génomique a été réalisé sur l'ensemble du génome avec des tests de neutralité et de différenciation, prenant en compte les effets démographiques. Cela a permis de proposer des régions génomiques porteuses de gènes candidats potentiellement impliqués dans le contournement de la résistance 7.

Chapitre 2
Impact des contournements de résistance sur le génome de *Melampsora larici-populina*

Chapitre 2 - Impact des contournements de résistance sur le génome de *Melampsora larici-populina*

1. Introduction

Cette étude porte sur l'analyse du polymorphisme génomique chez le champignon responsable de la rouille du peuplier, *M. larici-populina*, avec comme but l'identification des facteurs liés à la pathogenèse. Plus précisément, les objectifs sont d'identifier le niveau général et les caractéristiques du polymorphisme présent chez la rouille du peuplier et de le corréler aux profils de virulences des isolats considérés.

Pour cela, 15 isolats de *Melampsora larici-populina* ont été sélectionnés pour leurs profils de virulence distincts qui ont été testés par un pathotypage sur huit cultivars de peupliers portant chacun une et une seule résistance (de R1 à R8). Cela nous a permis de déterminer le profil de virulence correspondant (de Vir1 à Vir8) en fonction de la capacité à contourner ces résistances. Le génome de ces 15 isolats a été séquencé par la technologie Illumina HiSeq2000. Les séquences ainsi obtenues ont été alignées sur les 462 scaffolds de la version 1 du génome de référence 98AG31 (Duplessis et al., 2011a).

L'ensemble des variants (SNP et indels) ont été détectés, nous permettant de documenter un niveau de polymorphisme très élevé. Une première partie de l'étude a consisté en l'analyse de la profondeur de séquençage des différents scaffolds afin d'identifier des régions absentes chez certains des 15 isolats.

Nous nous sommes ensuite concentrés sur les gènes et plus particulièrement ceux codant pour des protéines sécrétées représentant des effecteurs candidats (Lowe and Howlett, 2012). Une étude d'enrichissement en catégories KOG (Eukaryotic Orthologous Group, un système d'annotation fonctionnelle des protéines) des gènes a été réalisée. Les taux de polymorphismes synonymes et non-synonymes ont ensuite été été calculés sur l'ensemble des gènes et comparés entre gènes codant des protéines sécrétées et les autres. Les gènes codant pour les protéines sécrétées présentant le plus fort excès de variabilité non-synonyme ont été corrélés aux profils de virulence des différents isolats nous permettant d'identifier des effecteurs candidats potentiellement impliqués dans le phénotype Vir7.

Il s'agit de la première étude mesurant la variabilité génomique (obtenue à partir de données de séquences à haut débit) au sein d'isolats de *M. larici-populina*, et elle ouvre la voie à des approches de génomique des populations plus exhaustives (voir chapitre 4).

Ce chapitre se présente sous la forme d'un article publié dans la revue *Frontiers in Plant Science* (doi: 10.3389/fpls.2014.00450)

2. Article n°1 : Patterns of genomic variation in the poplar rust fungus *Melampsora larici- populina* identify pathogenesis-related factors

Antoine Persoons[1,2], Emmanuelle Morin[1,2], Christine Delaruelle[1,2], Thibaut Payen[1,2], Fabien Halkett[1,2], Pascal Frey[1,2], Stéphane De Mita[1,2] and Sébastien Duplessis[1,2]

[1]INRA, Unité Mixte de Recherche 1136 INRA/Université de Lorraine, Interactions Arbres/Microorganismes, 54280 Champenoux, France
[2]Université de Lorraine, Unité Mixte de Recherche 1136 INRA/Université de Lorraine, 54506 Vandoeuvre-lès-Nancy Cedex, France

*Corresponding author: Dr. Sébastien Duplessis
INRA, Unité Mixte de Recherche 1136 INRA/Université de Lorraine, Interactions Arbres/Microorganismes, 54280 Champenoux, France
e-mail, duplessi@nancy.inra.fr

Abstract

Melampsora larici-populina is a fungal pathogen responsible for foliar rust disease on poplar trees, which causes damage to forest plantations worldwide, particularly in northern Europe. The reference genome of the isolate 98AG31 was previously sequenced using a whole genome shotgun strategy, revealing a large genome of 101 megabases containing 16,399 predicted genes, which included secreted protein genes representing poplar rust candidate effectors. In the present study, the genomes of 15 isolates collected over the past 20 years throughout the French territory, representing distinct virulence profiles, were characterized by massively parallel sequencing to assess genetic variation in the poplar rust fungus. Comparison to the reference genome revealed striking structural variations. Analysis of coverage and sequencing depth identified large missing regions between isolates related to the mating type loci. More than 611,824 single-nucleotide polymorphism (SNP) positions were uncovered overall, indicating a remarkable level of polymorphism. Based on the accumulation of non-synonymous substitutions in coding sequences and the relative frequencies of synonymous and non-synonymous polymorphisms (i.e., PN /PS), we identify candidate genes that may be involved in fungal pathogenesis. Correlation between non-synonymous SNPs in genes encoding secreted proteins (SPs) and pathotypes of the studied isolates revealed candidate genes potentially related to virulences 1, 6, and 8 of the poplar rust fungus.

INTRODUCTION

Worldwide, *Melampsora* spp. (Basidiomycota, Pucciniales) are the most devastating pathogens of poplars (Steenackers et al., 1996), and *Melampsora larici-populina* is a major threat in European poplar plantations (Pinon and Frey, 2005). The poplar rust fungus has a complex life cycle with five different types of spores that develop on two distinct host plants: *Populus*, on which it performs several asexual reproduction cycles during summer and autumn, and *Larix* spp., on which it performs a single sexual reproduction cycle once a year in spring. Poplars are particularly susceptible to *M. larici-populina* mostly because of their intensive monoclonal cultivation over several decades (Gérard et al., 2006). Until now eight qualitative resistances (*R1* to *R8*) have been deployed in plantations and each has been overcome by *M. larici-populina*. The most damaging resistance breakdown occurred in 1994 when the resistance *R7* was overcome and led to the invasion of France by virulent 7 *M. larici-populina* isolates (Xhaard et al., 2011). In accordance with the gene-for-gene relationship (Flor, 1971), *M. larici-populina* isolates which successfully infect resistant poplar possess the corresponding virulence factors (*i.e.* Vir1 to Vir8) determined at an avirulence locus. Up to now, none of the poplar R genes, nor the poplar rust virulence genes have been characterized (Hacquard et al., 2011).

Pathogenicity factors, i.e. effectors, contribute to the success of pathogen infection. Their recognition by cytoplasmic plant *R* receptors leads to a rapid and strong defense reaction through specific signalling cascades and expression of defense-related genes that stop pathogen growth, notably through the expression of a localized hypersensitive response at infection site (Dodds and Rathgen, 2010; Win et al., 2012). Most effectors described to date in rust fungi correspond to avirulence factors such as AvrL567, AvrP4, AvrP123 and AvrM of the flax rust fungus *Melampsora lini* (Ravensdale et al., 2011) and PGTAUSPE-10-1, a candidate AvrSr22 factor of the wheat stem rust *Puccinia graminis* f. sp. *tritici* (Upadhyaya et al., 2014), but their role in pathogenesis remain unknown. Another effector, the Rust Transferred Protein 1 (RTP1) from the bean rust fungus *Uromyces fabae*, forms fibrils in the extrahaustorial matrix and is transferred from haustoria into infected host cells, and may have protease inhibitory function (Kemen et al., 2005 ; Kemen et al., 2013; Pretsch et al., 2013). So far, only a handful of fungal candidate effectors have been fully characterized (Stergiopoulos and de wit, 2009; Tyler and Rouxel, 2012; Giraldo and Valent, 2013). Fungal effectors share several features, which are not exclusive, i.e. most have a N-terminal secretion signal, enrichment in cysteine residues and a lack of functional homology in databases and present a small size. Such features have been widely used to determine sets of candidate effectors in the predicted proteome of fungal pathogens for which a reference genome has been sequenced (Lowe and Howlett, 2012; Duplessis et al., 2014a).

Host immunity escape by pathogens is frequently mediated by deletion or mutations in

effector genes, which often show elevated levels of non-synonymous polymorphism as a result of their antagonistic co-evolution with the host (Stukenbrock and McDonald, 2009). The relative abundance of non-synonymous and synonymous polymorphisms (P_N and P_S) measures the direct effect of positive selection that tends to remove deleterious non-synonymous variants in coding sequences. When considered at the interspecific level, the rates of non-synonymous and synonymous substitutions (termed d*N* and d*S*, respectively) can be assessed to contrast patterns of variation between species (Stukenbrock and Bataillon, 2012). Such approaches have been applied at the genome scale to detect sets of candidate effectors in oomycetes and fungi (Raffaele and Kamoun, 2012; Cantu et al., 2013; Stukenbrock, 2013; Stergiopoulos et al., 2013). Evidence of positive selection was reported in avirulence genes of rust fungi at the intraspecific (AvrL567, Dodds et al., 2004; AvrP4 and AvrP123, Barrett et al., 2009) or interspecific levels (AvrP4, Van der Merwe et al., 2009). Genome-scale approaches were also used with sets of candidate effectors at the intraspecific level in *Puccinia striiformis* f. sp. *tritici* (Cantu et al., 2013) or by considering clusters of paralogous genes in the genome of *M. larici-populina* (Hacquard et al., 2012).

Genomics is becoming a method of choice to identify new candidate effectors, particularly in obligate biotrophs where functional approaches are impeded. Only a handful of rust fungi genomes are available (Duplessis et al., 2011a; Cantu et al., 2011; Zheng et al., 2013; Cantu et al., 2013; Nemri et al., 2014). In these, repertoires of candidate effectors corresponding to small secreted proteins (SSPs) have been defined (Hacquard et al., 2012; Saunders et al., 2012; Cantu et al., 2013; Zheng et al., 2013; Nemri et al., 2014). The poplar-poplar rust pathosystem is a model in forest pathology because it is one of the few pathosystems for which both the host and pathogen genomes are available (Tuskan et al., 2006; Duplessis et al., 2011a). *M. larici-populina* has a remarkably large diploid genome of 101 Mb enriched in repetitive and transposable elements (TE), a common feature of rust fungi genomes. There is a striking number of 16,399 predicted genes in the poplar rust genome, another feature shared with other rust fungi (Duplessis et al., 2014b). Among genes encoding secreted proteins (SPs), a set of 1,184 SSP genes showing typical features of pathogen effectors was uncovered; most of these are cysteine-rich, belong to multigene families and are lineage specific (Duplessis et al., 2011a; Hacquard et al., 2012). In order to prioritize functional analysis of such candidates, other features were searched including specific expression during the interaction with the poplar host (Duplessis et al., 2011b), presence of conserved motifs in proteins, and gene families exhibiting evidences of positive selection by considering a classification into clusters of paralogous genes (CPG) (Joly et al., 2010; Hacquard et al., 2012). Another way to identify promising effectors is to study gene polymorphism at the intraspecific or interspecific level, as has been performed in *M. lini* (Ravensdale et al., 2011).

In the present study, we report on the genome sequencing of 15 *M. larici-populina* isolates

and their comparison to the reference genome of isolate 98AG31 (Duplessis et al., 2011a) in order to identify patterns of genomic variations that may relate to fungal pathogenesis. Genes that accumulate intraspecific polymorphism in their coding sequence as well as in their non-coding upstream regions were scrutinized, thus providing a new filter to prioritize candidate effectors of interest.

MATERIALS AND METHODS

Fungal material and DNA preparation

Isolates were selected in a laboratory collection (Frey P., INRA Nancy, Champenoux, France) in order to maximize historical and geographical repartitions and virulence profiles (Table 1). All isolates were genotyped based on 25 microsatellite loci, confirming the purity of each isolate and the absence of clones among the selected isolates. Phenotypes of all isolates (i.e. combination of virulences) were confirmed in triplicate on eight poplar cultivars each carrying a single resistance (R1 to R8) to *M. larici-populina* (Table 1) and on the universal clone 'Robusta', as a positive control. To ensure their purity and to avoid potential clones within the selected isolates, genotyping was performed using the 25 microsatellite markers (Xhaard et al., 2011). Urediniospores of each isolates were multiplied on '*Robusta'* detached leaves to obtain enough material for genomic DNA isolation.

DNA isolation

A total of 100-300 mg of urediniospores were used for DNA isolation using a CTAB method. Spores were crushed using a Retsch Tissue Lyser (Qiagen, Courtaboeuf, France) at a frequency of 30Hz for 1 min. Broken spores were resuspended in CTAB buffer (Tris 0.1 M, NaCl 1.43 M, EDTA 0.02 M, CTAB 0.02 M) and heated at 65°C for 30 minutes. The suspension was subjected to centrifugation at 8,000 rpm at room temperature for 5 min to pellet spore debris. Supernatant was gently mixed with an equal volume of phenol:chloroform:isoamyl alcohol (50:48:2; Euromedex, Souffelweyersheim, France) and centrifuged at 8,000 rpm at room temperature for 10 min. The aqueous phase was recovered, gently mixed with an equal volume of chloroform and centrifuged at 8000 rpm at room temperature for 10 min. The aqueous phase was subjected to RNA digestion with RNaseA at 10 µM (Fermentas, Saint-Remy-lès-chevreuses, France) at 37°C for 30 min. A final extraction with an equal volume of chloroform was realized followed by centrifugation at 8000 rpm at room temperature for 10 min. The recovered aqueous phase was then subjected to isopropanol (0.75 of final volume) precipitation, followed by centrifugation at 14,000 rpm at 4°C for 30 min.

Isolate	Year	Location	Latitude, Longitude	Host	Pathotype
93ID6	1993	Champenoux (NE France)	N 48° 45' 02", E 06° 20' 20"	*P. x euramericana* 'I45-51'	3-4
02Y5	2002	Charrey-sur-Saône (NE France)	N 47° 05' 18", E 05° 09' 11"	*P. x euramericana* 'Robusta'	2-3-4-7-8
09BS12	2009	Mirabeau (SE France)	N 43° 41' 29", E 05° 40' 21"	*P. nigra*	4-6
94ZZ15	1994	Saulchoy (N France)	N 50° 21', E 1°50'	*P. x euramericana* 'Luisa Avanzo'	3-4-5-7
94ZZ20	1994	Nogent-sur-Vernisson (Central France)	N 47° 50', E 2° 45'	*P. x interamericana* 'Boelare'	3-4-7
08EA47	2008	Prelles (SE France)	N 44° 51' 00", E 06° 34' 47"	*P. nigra*	2-4
95XD10	1995	Rogécourt (N France)	N 49° 39', E 3° 25'	*P. x euramericana* 'Flevo'	3-4-5-7
08EA20	2008	Prelles (SE France)	N 44° 51' 00", E 06° 34' 47"	*P. nigra*	4
08EA77	2008	Prelles (SE France)	N 44° 51' 00", E 06° 34' 47"	*P. nigra*	4-6
97CF1	1997	Champenoux (NE France)	N 48° 45' 02", E 06° 20' 20"	*P. x interamericana* 'Hoogvorst'	3-4-7
08KE26	2008	Mirabeau (SE France)	N 43° 41' 29", E 05° 40' 21"	*P. nigra*	4
9683B13	1996	Orléans (Central France)	N 47° 49' 39", E 1° 54' 40"	*P. x interamericana* '83B13'	1-3-4-5-6-7
98AG31	1998	Moy-de-l'Aisne (N France)	N 49° 45', E 3° 21'	*P. x interamericana* 'Beaupré'	3-4-7
93JE3	1993	Champenoux (NE France)	N 48° 45' 02", E 06° 20' 20"	*P. x euramericana* 'Blanc du Poitou'	2-4
98AR1	1998	Geraardsbergen (Flanders, Belgium)	N 50° 45', E 3° 52'	*P. x interamericana* 'B71085/A1'	1-3-4-5-7-8

Table 1: Summary of *Melampsora larici-populina* isolates. Isolate name, year and location of sampling are indicated. Host indicates the poplar species/cultivar on which the isolate was sampled. The pathotype profile (combination of virulences) was confirmed in triplicate by inoculation on a differential set of poplar cultivars carrying the eight known resistances to *M. larici-populina*.

DNA pellet was washed twice with 70%, then absolute ethanol, each followed by centrifugation at 14,000 rpm at 4°C for 10 min. The DNA pellet was finally dried under a hood for 20 min and resuspended in 1X Tris EDTA. Quality and quantity of recovered high molecular weight DNA was assessed by electrophoresis on agarose gel, by spectrophotometry (Nanodrop, Saint-Remy-lès-Chevreuse, France) and with the QuBit (Life Technologie, Villebon-sur-Yvette, France) fluorometric quantitation system.

Genome re-sequencing

For all isolates, except 98AR1, genomic DNA libraries were prepared using TruSeq DNA sample preparation kit (v3) followed by paired-end 100 nt massively parallel sequencing on Illumina HiSEQ2000 by Integragen (Evry, France). Briefly, 3 µg of each genomic DNA were fragmented by sonication and purified to yield fragments of 400-500 nt. Paired-end adapter oligonucleotides from Illumina were ligated on repaired A-tailed DNA fragments, then purified and enriched by PCR cycles. Each library was quantified by qPCR and sequenced on Illumina HiSeq2000 platform as paired-end 100 nt reads. Image analysis and base calling were performed using Illumina Real Time Analysis (RTA 1.13.48.0) pipeline with default parameters. Isolate 98AR1 genomic DNA was sequenced by a single read strategy of 75bases on Illumina Genome Analyzer II (Beckman Coulter Genomics, Grenoble, France).

Filtering and mapping of short reads

Adapter and quality filtering was carried out using CLC Genomics Workbench 6.5 (CLC bio, QIAgen, Aarhus, Denmark). For each batch of reads, 3 and 10 low quality terminal nucleotides were trimmed at the 5' and 3' ends, respectively. FASTQ files of trimmed sequences were used to proceed with mapping onto the 98AG31 reference genome available at the Joint Genome Institute (JGI; http://genome.jgi.doe.gov/programs/fungi/index.jsf; Duplessis et al., 2011a). The 462 scaffolds composing the reference genome were uploaded in CLC Genomics Workbench and the annotation was superimposed onto the scaffolds using the annotation plugin. The following parameters were applied for mapping: masking mode = no masking; mismatch cost = 2; insertion cost = 3; deletion cost = 3; length fraction = 1.0; similarity fraction = 0.95; global alignment = no; auto-detect paired distances = yes; non-specific match handling = map randomly. Sequencing data and assemblies were deposited at at the National Center for Biotechnology Information (NCBI) and the Short Reads Archive (Bioproject PRJNA251864 study SRA accession SRP042998). Coverage and sequencing depth values were extracted from the CLC stand-alone read mapping files and were further used to compare scaffolds of resequenced isolates. Sequencing depth and coverage on each scaffold were visually inspected using the CLC read tracks functions used for further detection of

structural variants.

Scaffold depth analysis and variants detection

Cross-comparison of average coverage and sequencing depth onto the 462 reference scaffolds was performed within and between isolates based on the CLC Genomics Workbench mapping outputs to detect the potential presence/absence of regions and the sequencing coverage or depth bias. In the case of missing regions or coverage bias, read mapping profiles and distribution of genes and TEs on the scaffolds were inspected manually. In these manual inspections, regions with high concentrations of ambiguous mappings were excluded from consideration, because of the possibility of artifactually divergent coverage. In parallel, the coverage analysis tool implemented in CLC Genomics Workbench (version 7.0) was used to detect regions within scaffolds showing significantly unexpected low or high coverage relative to the reference genome,according to a Poisson distribution of observed coverage in mapping positions (*p*-value threshold = 0.0001 and minimum length of the coverage region of 100 bp). Search for SP genes in the low-coverage regions was performed using an in-house Python script. Notably, this script was limited to detection of genes which laid entirely inside the corresponding region.

Single Nucleotide Variants (SNVs, i.e. Single Nucleotide Polymorphisms, SNPs), Multiple Nucleotide Variants (MNVs, i.e. successive SNVs), and small Insertion/Deletion variants (i.e. InDels) were detected in the genome of each isolate based on mapping outputs using the quality-based variant detection option of CLC Genomics Workbench (version 6.5.1). This option considers minimum quality levels and minimum coverage of bases where the variant is detected and in surrounding bases. The following parameters were considered: neighborhood radius = 5; maximum gap and mismatch count = 2; minimum neighborhood quality = 15; minimum central quality = 20; ignore non-specific matches = yes; ignore broken pairs = yes; minimum coverage = 10; minimum variant frequency = 35%; maximum expected alleles = 2; advanced = no; require presence in both forward and reverse reads = no; ignore variants in non-specific regions = no; genetic code = standard. Variant tables were generated for all isolates. Selection of synonymous and non-synonymous polymorphism in genes and variants in 1 Kb upstream regions of genes was performed using in-house Python scripts.

Sequence analysis

Gene and protein sequences and Gene Ontology (GO) and Eukaryotic Orthologous Group (KOG) functional annotations were retrieved from the *M. larici-populina* genome sequence on the Mycocosm website at the JGI (http://genome.jgi.doe.gov/programs/fungi/index.jsf). Homology searches were carried out using the Blastp algorithm (Altschul et al., 1997) against the non-

redundant database at the NCBI (March 2014). AvrP4 sequences from Van der Merwe et al. (2009) and Barrett et al. (2009) were retrieved from the NCBI and used for multiple alignments with members of the CPG5464 gene family previously identified in the *M. larici-populina* genome (Hacquard et al., 2012). Alignment with variants of the CPG5464 gene family retrieved in the *M. larici-populina* isolates was conducted using the program ClustalW (Thompson et al., 1994) and gaps were manually inserted to strictly align sites reported under positive selection in the above-mentioned articles, before generating conservation profiles on the WebLogo server (Crooks et al., 2004).

KOG enrichment analysis

KOG (Tatusov et al., 2003) annotation of each *M. larici-populina* gene was retrieved from the JGI genome website. Each gene was classified according to the KOG functional classification using custom Perl scripts. Over-represented KOG categories in a selected gene set were calculated relative to the global gene distribution in the genome. Fisher's exact test was used to determine significant differences in the distribution of genes by KOG categories between the selected gene set and all genes ($p < 0.05$).

P_N/P_S analysis

For each gene, an alignment was generated with a custom Python script based on the reference genome and gene annotations (gff files from the *M. larici-populina* JGI website) taking into account the SNP variants generated by CLC Genomics Workbench. Alignments interrupted by an early stop codon were excluded from the computation of synonymous and non-synonymous polymorphisms. Polymorphism index was computed for each gene using Egglib version 2.1.6 (De Mita and Siol, 2012). This Python library computes from an alignment the number of synonymous or non-synonymous sites either polymorphic or non-polymorphic. P_N/P_S is computed as the ratio of the number of synonymous over non-synonymous polymorphisms corrected by the number of synonymous and non-synonymous sites, respectively.

RESULTS

Sequencing efficiency

Genomes of 15 *M. larici-populina* isolates, including the 98AG31 reference isolate, were sequenced at a targeted depth of ~40X. A total of 64 billion bases were generated, corresponding to 25 to 63 million reads per genome. After filtering, the average read length was 84.4 nt. A number of length and similarity parameters were tested for maping reads onto the reference genome. Loose default

parameters tended to generate multiple mappings in repetitive sequences including large gene families, impinging on further call of variants in a given isolate (data not shown). Stringent parameters were retained (i.e. total length of the sequence showing a minimum of 95% similarity) for optimal mapping and subsequent variant calling. On average, 78% of the reads aligned to the 462 scaffolds of the reference genome (63 to 90%), and only one isolate had a lower percentage of mapped reads (isolate 9683B13, 40%). Examination of 1,000 randomly selected unmapped reads from genome 9683B13 showed contamination with bacterial sequences (68%; >30% *Pseudomonas* sp. and >10% *Stenotrophomonas maltophilia*, data not shown), so these sequences were discarded. Overall, this led to a sequencing depth average of 32X per genome (22X to 46X; Table 2). Overall coverage was between 90.7% and 96.3% for the 15 isolates. For all genomes sequenced with paired-end reads (that is, all except 98AR1), the number of broken paired reads was relatively moderate (<11% and average of 9%).

Coverage and sequencing depth analysis

Cross-comparison of mapping outputs identified a bias of average coverage and sequencing depth among the 462 reference scaffolds within and between isolates. For instance, several scaffolds systematically showed very high (> 100X) or low (< 1X) depths in all sequenced isolates, and others showed discrepancies for a given scaffold between different isolates. Such situations were manually inspected and led to the survey of 151 scaffolds (representing about 10% of the genome sequence) for which the mapping depth profile and the presence of genes along the scaffolds were recorded (Supporting Table 1). Notably, scaffold 484 showed a systematic high depth > 1,000X. Four mitochondrial scaffolds were previously identified and removed from the poplar rust genome assembly (Duplessis et al., 2011a). Mapping of Illumina reads from the 15 isolates onto these four scaffolds showed much higher depth than the average observed for other scaffolds (178X to 1,211X, data not shown). Inspection of scaffold 484 indicated that it is most likely a portion of the mitochondrial genome. Indeed, this 5.4 Kb scaffold bears two genes showing high homology to two mitochondrial genes (ATP synthase F0 subunit and NADH dehydrogenase subunit).

For other scaffolds with systematic high coverage and sequencing depth biases, major differences are explained by missing regions in one or several isolates. Such scaffolds were marked by no mapping support for the entire scaffold, or for some regions of the scaffold at the same positions in a given subset of isolates (i.e. probable large deletions or highly variable loci). For instance, the 319 Kb scaffold 90 showed either a similar depth along the scaffold in reference isolate 98AG31 and five other isolates (pattern A; Figure 1), or the absence of regions at the same positions for two patterns, each grouping different isolates (patterns B and C; Figure 1). Pattern C exhibited an overall low sequencing depth ranging from 3.5X to 5.8X, that mostly corresponds to

Isolate	Total reads number	Mapped reads	% mapped reads	Broken pairs	Average read length	Sequencing depth
93ID6	3,594,455,577	2,656,764,147	73.9	226,296,523	84.4	26.3
02Y5	3,691,995,994	3,218,997,193	87.2	269,105,383	85.4	31.8
09BS12	6,230,429,688	4,717,557,005	75.7	479,213,815	84.2	46.6
94ZZ15	3,653,741,644	3,290,238,877	90.1	278,395,986	85.3	32.5
94ZZ20	3,387,309,786	3,045,158,939	89.9	253,470,401	85.2	30.1
08EA47	4,659,300,813	3,460,505,640	74.3	352,258,523	83.3	34.2
95XD10	4,701,407,950	3,993,529,488	84.9	396,812,163	83.7	39.5
08EA20	4,829,802,826	3,034,419,164	62.8	290,972,918	83.2	30.0
08EA77	4,259,571,919	3,840,082,037	90.2	340,127,111	84.7	38.0
97CF1	3,570,560,916	3,083,826,749	86.4	270,564,864	84.5	30.5
08KE26	5,407,393,523	4,626,803,739	85.6	434,085,871	85.0	45.8
9683B13	6,378,404,736	2,537,206,558	39.8	223,243,679	83.1	25.1
98AG31	2,779,485,081	2,529,716,573	91.0	218,294,868	85.2	25.0
93JE3	4,310,048,066	2,796,054,256	64.9	258,175,513	84.1	27.6
98AR1	2,562,143,464	2,227,530,892	86.9	na	76.0	22.0

Table 2: General mapping information for the 15 *Melampsora larici-populina* isolates. Illumina reads of each genome were mapped onto the 98AG31 JGI reference genome. na, not applicable.

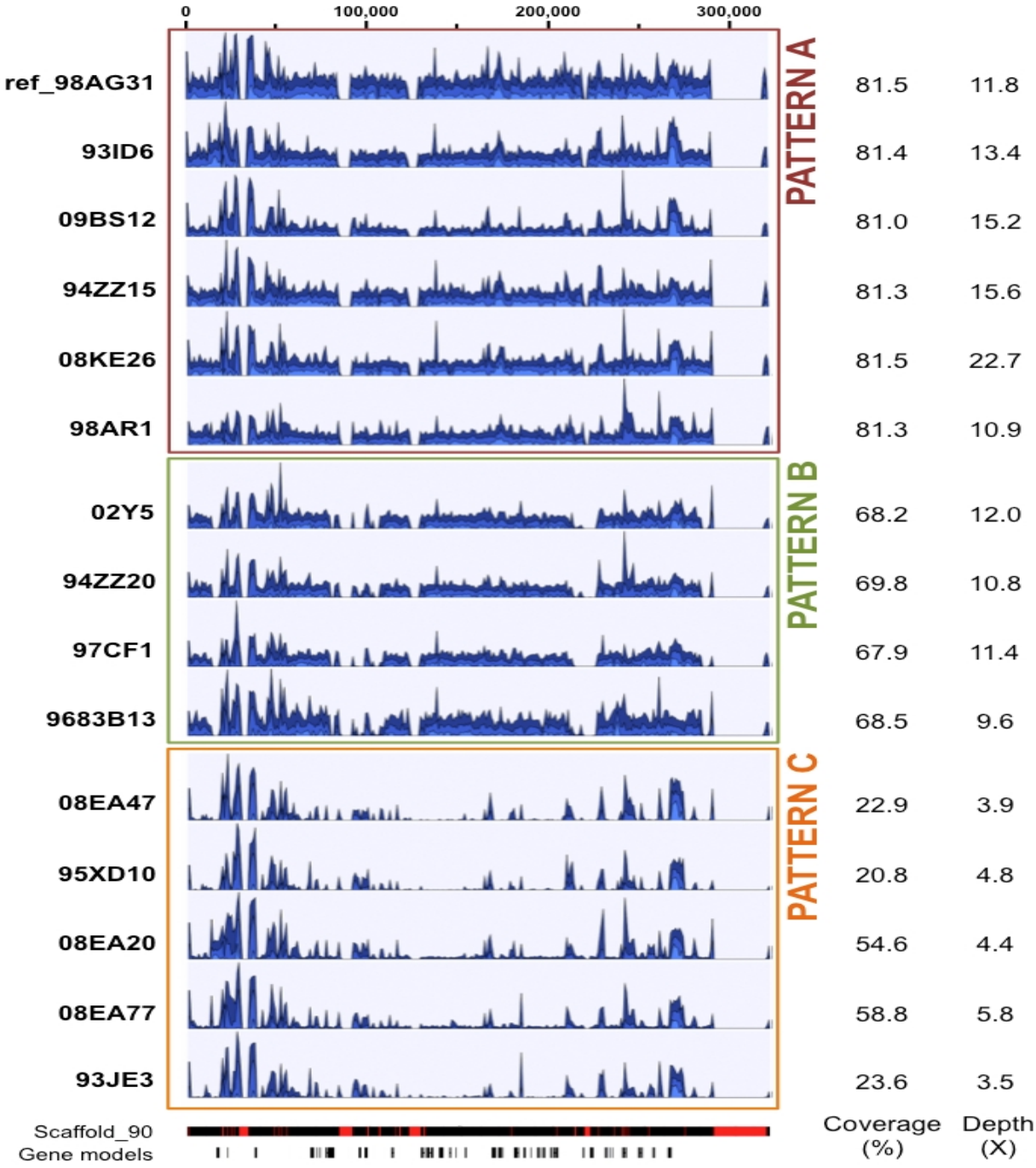


Figure 1: Patterns of sequencing depth along scaffold 90 in 15 *Melampsora larici-populina* isolates. Illumina reads from 15 isolates were mapped onto the 98AG31 reference genome. Scaffold 90 is presented here to illustrate distinct patterns of sequencing depth between groups of isolates: pattern A (red box) with coverage and sequencing depth similar to 98AG31, pattern B (green box) presenting four regions of lower coverage, and pattern C (orange box) with overall reduced coverage. Graphical outputs in blue represent the local sequencing depth along scaffold 90, normalized to the maximum depth measured in each isolate. Average coverage and sequencing depth are detailed for each isolate on the right. The bars below represent scaffold 90 from JGI reference genome website (red blocks indicate gaps) and predicted gene models (38 in total). Scale in nucleotides is presented at the top. The total scaffold length is 319,043 bp.

repetitive elements regions marked by peaks of high depth similar to those present in patterns A and B. This indicates that the missing regions were not related to sequencing depth (Figure 1). For pattern C with the longest missing regions, a total of 38 genes were not supported by reads, including 4 pheromone genes related to mating type in the poplar rust fungus. Despite a generally similar profile and sequencing depths within pattern C, isolates 08EA20 and 08EA77 showed a higher coverage (54.6 and 58.8%, respectively) than the other three isolates (20.8, 22.9 and 23.6%). This is explained by a light and continuous depth in the central region of the scaffold that was totally absent in the other isolates (Figure 1). In isolate 08EA20 two genes located at 16-17 Kb (hypothetical protein) and 22-23 Kb (chitinase) were present. In isolate 08EA77, only the chitinase encoding gene was present, whereas these two genes were missing in the other three isolates of pattern C. Assembling unmapped reads from isolates exhibiting pattern C onto the 38 missing genes using loose similarity parameters retrieved only highly divergent and/or partial sequences (data not shown). Because of the presence of pheromone genes on scaffold 90, we looked at previously described mating type loci in the *M. larici-populina* genome (Duplessis et al., 2011a). A missing region containing a pheromone gene and a STE3 pheromone receptor gene was also observed in scaffold 172 for the isolates with pattern C. This prompted us to examine the homeodomain locus, composed of the genes HD1 and HD2. The five isolates that exhibited missing regions in scaffolds 90 and 172 also presented a missing region at the homeodomain locus in the scaffold 35. Using the homeodomain loci and pheromone/receptor loci genes as baits, divergent alleles were identified for *M. larici-populina HD1*, *HD2* and some pheromone genes in the unmapped reads of these isolates (data not shown).

A total of 212 genes lie in the missing regions of the surveyed scaffolds, including 12 SP genes in 7 scaffolds (Supporting Table 1). We therefore conducted a systematic analysis of regions of 100 bp or more showing coverage differences using the CLC coverage analysis tool, in order to detect possible deletions or amplifications. In total, 18,564 to 81,325 regions with significantly high/low coverage differences relative to the 98AG31 reference genome were identified in the 14 isolates (Supporting Table 2). Search for SP genes within these regions revealed that between 12 (9683B13) and 59 (95XD10) SP genes are in low coverage regions, indicating a possible deletion compared to isolate 98AG31. However, we could not find any correlation between a probable SP gene deletion and the pathotypes of the isolates, i.e., the absence of a SP gene explaining virulences 1, 2, 5, 6 or 8 (98AG31 reference isolate being virulence 3, 4, 7).

Polymorphism and insertion/deletion detection

In order to assess polymorphism in the 15 isolates, variants (SNVs/SNPs, MNVs and InDels) were recorded using the CLC Genomics Workbench program. The 98AG31 reference genome had been

sequenced at a 6.9X sequencing depth from dikaryotic urediniospores by Sanger sequencing, following a whole-genome shotgun strategy. Therefore the 462 scaffolds represent a chimeric version of the genome combining the two haplotypes (Duplessis et al., 2011a). Resequencing by Illumina at a sequencing depth of 25X identified a total of 93,189 variants including 86,877 SNPs, 1,741 MNVs, 2,945 insertions and 1,626 deletions in isolate 98AG31 (representing 96,099 bases; Table 3), which is in close range with the 88,083 SNPs recorded by Sanger sequencing. However, only 40,001 SNPs from the initial assembly were confirmed, highlighting differences due to the sequencing approaches. An average of 163,477 variants (including 152,936 SNPs) representing 168,708 bases was found in the 14 other isolates mapped onto the reference genome, representing a larger number of polymorphic sites at the inter-individual level (0.17% of the genome; 1.51 SNPs/Kb). When the 15 genomes were considered together, 11,683 SNPs were conserved, whereas in total 611,824 unique SNPs were found. The variant caller implemented in CLC allowed the determination of the zygosity of nucleotides at the polymorphic sites. The heterozygosity rate was 0.45-0.55 in 12 isolates, whereas it was lower in 09BS12 and 08KE26 (0.35 and 0.37, respectively) and higher in 98AG31 (0.85). The latter is as expected, as it was the reference genome to which reads were mapped (Table 3). For all genomes, the ratio of transition over transversion mutations was 2.31± 0.11 (Table 4), which is similar in range to previous observations in rust fungi (Cantu et al., 2013). Individually, all isolates except the reference 98AG31 showed similar numbers of SNPs, MNVs and InDels (Table 3), indicating a homogeneous polymorphism rate at the intraspecific level. Polymorphic sites residing within coding DNA sequences (CDS) were more closely scrutinized and represented 20% of the SNPs, 17% of the MNVs, and 5% of deletions and 5% of insertions in InDels. These proportions were rather similar in the different isolates (Table 4). In total, more SNPs were present in exons than in introns (average 30,077 ± 3,893 SD and 14,982 ± 1,871 SD, respectively; Table 4), but when exon and intron size were accounted for, introns tended to accumulate more SNPs than the coding sequences (data not shown).

Highly variable genes

Synonymous and non-synonymous polymorphisms within the 15 isolates were inspected in the gene complement of *M. larici-populina*, considering only SNPs that were represented in most of the observed variants (90%). Both homozygous and heterozygous SNPs were considered. For cross-comparison of SNPs between isolates, non-redundant SNPs (i.e., nucleotides in the reference isolate presenting polymorphism in at least one other isolate) were considered. Overall, a very large portion of the genes (89%) was marked at least by one SNP, and 5,332 and 10 genes exhibited more than 10 and 100 SNPs, respectively (Supporting Table 3). A total of 1,089 genes in the 15 isolates had more than 10 non-synonymous SNPs in CDS, the maximum number being 66 (proteinID 66139).

Isolate	Zygosity	Variant types	Total					
	Homozygous	Heterozygous	Deletion	Insertion	MNVs	SNVs	Variants	Nucleotides
93ID6	84,849	88,855	3534	4198	3302	162,670	173,704	179,274
02Y5	76,511	95,418	3514	4399	3348	160,668	171,929	177,658
09BS12	91,934	54,500	3170	4020	2835	136,409	146,434	151,298
94ZZ15	84,155	82,478	3485	4287	3160	155,701	166,633	172,001
94ZZ20	80,851	80,541	3385	4085	3002	150,920	161,392	166,613
08EA47	85,423	75,527	3435	4158	3026	150,331	160,950	166,117
95XD10	68,735	87,520	2909	3554	2903	146,889	156,255	160,886
08EA20	90,268	91,000	3723	4354	3469	169,722	181,268	187,146
08EA77	89,765	83,569	3599	4275	3222	162,238	173,334	178,887
97CF1	75,954	76,585	3061	3902	2958	142,618	152,539	157,525
08KE26	102,244	55,022	3578	4268	3100	146,320	157,266	162,670
9683B13	70,974	82,208	3004	3708	2866	143,604	153,182	157,967
98AG31	14,219	78,970	1626	2945	1741	86,877	93,189	96,099
93JE3	91,933	75,793	3277	3938	3182	157,329	167,726	172,951
98AR1	77,267	88,799	3315	3932	3130	155,689	166,066	170,921

Table 3: Genomic variants identified in 15 *Melampsora larici-populina* isolates by mapping onto the 98AG31 JGI reference genome. MNV, Multiple Nucleotide Variant; SNV, Single Nucleotide Variant (i.e. Single Nucleotide Polymorphism).

	SNPs				% polymorphism in CDS				
Isolate	**Tr/Tv**	**SNPs in exon**	**SNPs in intron**	**SNPs intergenic**	**Non-synonymous SNP**	**Deletion**	**Insertion**	**MNV**	**SNV**
93ID6	2.30	33,428	16,489	112,753	15,950	5.0	5.7	18.3	20.5
02Y5	2.30	32,904	16,325	111,439	15,553	5.5	5.4	15.9	20.5
09BS12	2.34	26,086	13,365	96,958	12,905	5.2	5.6	16.8	19.1
94ZZ15	2.29	31,938	16,056	107,707	15,252	6.2	5.7	17.2	20.5
94ZZ20	2.29	31,035	15,461	104,424	14,859	5.2	5.2	17.7	20.6
08EA47	2.30	29,848	14,986	105,497	14,493	5.3	5.5	17.3	19.9
95XD10	2.42	29,932	14,817	101,940	15,950	4.5	4.9	14.8	20.4
08EA20	2.30	35,069	17,230	117,423	16,911	5.3	5.4	18.0	20.7
08EA77	2.35	32,152	16,383	113,703	15,653	5.4	5.4	17.1	19.8
97CF1	2.33	29,566	14,649	98,403	14,218	5.7	6.1	17.9	20.7
08KE26	2.36	27,137	13,886	105,297	13,862	5.3	5.6	16.5	18.5
9683B13	2.33	29,776	14,719	99,109	14,442	5.1	5.5	17.7	20.7
98AG31	2.27	18,749	9335	58,793	8825	6.6	5.8	19.9	21.6
93JE3	2.36	32,155	15,684	109,490	15,651	4.9	5.4	18.0	20.4
98AR1	2.21	31,389	15,352	108,948	14,441	4.7	5.2	17.0	20.2

Table 4: Analysis of polymorphism in 15 *Melampsora larici-populina* isolates. CDS, Coding DNA sequence. Tr/Tv, rate of transition to transversion; MNV, Multiple Nucleotide Variants; SNV/SNP, Single Nucleotide Variant/Polymorphism.

JGI Protein ID[a]	Protein length	Transcript length	SNP	NS	Annotation	GO ID[a]	KOG ID[a]
66139	5273	15819	227	66	AAA ATPase, sigma 54 factor	0003677	1808
84101	1325	3975	95	57	Unknown protein	No hit	No hit
93626	1737	5211	82	54	Unknown protein	No hit	No hit
62079	1821	5463	73	47	Rhodopsin-like GPCR superfamily	0001584	No hit
106057	2195	6585	136	45	Unknown protein	No hit	No hit
92944	1135	3405	71	45	Unknown protein	No hit	No hit
95670	893	2679	87	45	Serine/threonine protein kinase	No hit	1187
66458	929	2787	55	45	Chromatin remodeling complex WSTF-ISWI	No hit	1245
70222	1542	4626	73	44	ATP-dependent DNA helicase	No hit	0351
101154	1470	4410	79	44	Unknown protein	No hit	No hit
114610	948	2844	91	41	Unknown protein	No hit	No hit
85441	1256	3768	56	40	Molecular chaperone (DnaJ superfamily)	No hit	0714
92226	1393	4179	59	38	Unknown protein	No hit	No hit
67208	1203	3609	76	37	Transcription regulator XNP/ATRX	No hit	1015
108793	931	2793	54	37	Unknown protein	No hit	No hit
96388	1344	4032	63	36	Unknown protein	No hit	No hit
108574	2851	8553	114	35	Unknown protein	No hit	No hit
91870	1131	3393	72	35	MHCK/EF2 kinase	0004674	3614
118268	1649	4947	108	34	Serine/threonine dehydratase	0006520	No hit
68278	1507	4521	54	34	FOG: Immunoglobin and related proteins	No hit	4475
65221	568	1704	44	34	Unknown protein	No hit	No hit
91258	771	2313	51	33	Nucleolar GTPase/ATPase p130	No hit	2992
88323	575	1725	55	33	Unknown protein	No hit	No hit
60895	698	2094	58	33	Nucleolar GTPase/ATPase p130	No hit	2992
84177	639	1917	57	33	Unknown protein	No hit	No hit
92190	551	1653	52	33	C-5 cytosine-specific DNA methylase	0006306	No hit
101664	1102	3306	63	32	Unknown protein	No hit	No hit
95815	1486	4458	46	32	Chromatin remodeling complex WSTF-ISWI	No hit	1245
107058	720	2160	51	32	Unknown protein	No hit	No hit
64441	1107	3321	45	31	Unknown protein	No hit	No hit

[a] Protein ID number, Eukaryotic Orthologous Group (KOG) and Gene Ontology (GO) annotations were retrieved from the 98AG31 reference genome at the Joint Genome Institute Mycocosm website (http://genome.jgi.doe.gov/programs/fungi/index.jsf)

Table 5: Top 30 genes accumulating non-synonymous (NS) Single Nucleotide Polymorphism (SNP).

Table 5 presents the top 30 genes with the highest number of non-synonymous SNPs over the 15 genomes, with 21.1 SNPs/Kb and 12 non-synonymous SNPs/Kb on average. Homology searches by Blastp against the NCBI nr protein database indicated a putative function for seven of the genes, six of which are associated with predicted nuclear activity. In total, 12 genes had GO and/or KOG annotations, and the majority encode predicted proteins of unknown function. A functional KOG analysis of the 4,142 genes exhibiting ≥ 5 non-synonymous SNPs revealed significant enrichment for gene categories related to chromatin structure and dynamics; cell cycle control, cell division and chromosome partitioning; nuclear structure; defense mechanisms and extracellular structures (Figure 2). SNPs were also inspected in the 1 Kb upstream regions of CDS, where they may impact transcription. Most genes also had at least one polymorphic site in their 1 Kb upstream regions (89%) and 2,554 genes each had more than 10 SNPs in these regions (Supporting Table 3). Half of the 30 genes with the highest number of SNPs had an annotation in various cellular categories including two SSP genes, the other half corresponded to genes encoding predicted proteins of unknown function (Supporting Table 4).

Highly variable secreted protein encoding genes

A set of 1,184 SSP-encoding genes representing candidate poplar rust effectors was previously reported (Hacquard et al., 2012). Because larger effectors were also described (e.g. flax rust AvrM; Ravensdale et al., 2011), we decided to place a particular focus on secreted protein encoding genes as possible candidate effectors (i.e. a total of 2,050 SPs identified by automatic annotation, including the 1,184 SSPs). We further distinguish SSPs from SPs as SSP genes were manually annotated in the *M. larici-populina* genome (Hacquard et al., 2012). Overall, a very large portion of the SP genes (89%) was marked by at least one SNP and 586 exhibited 10 SNPs or more (Supporting Table 5). A total of 386 and 119 genes had more than 5 and 10 non-synonymous SNPs (i.e. in CDS), respectively (maximum=45 non-synonymous SNPs; proteinID 66458). Table 6 presents the top 30 SP genes with the highest numbers of non-synonymous SNPs/Kb, of which 24 are SSP genes. Only six SPs showed homology to other fungal proteins, including an *M. lini* avirulence factor AvrP4, a metallopeptidase, and a pleckstrin homology-like domain involved in binding to interacting protein partners. Rates of synonymous (P_S) and non-synonymous (P_N) substitutions were calculated for all genes with the EggLib package (Supporting Table 3) and SP genes were more particularly scrutinized. The P_N/P_S rate could be measured for 14,052 genes, while 1,073 genes had a mutation generating a stop codon in the sequence and were excluded. P_N/P_S showed similar distributions between SP genes and other genes (Figure 3) and the highest P_N/P_S (4.9) was found for a gene encoding a hypothetical protein (ProteinID_70080; Supporting Table 3). In SP genes, the highest P_N/P_S was 2.47 and corresponds to a SSP of 200 amino acids with three

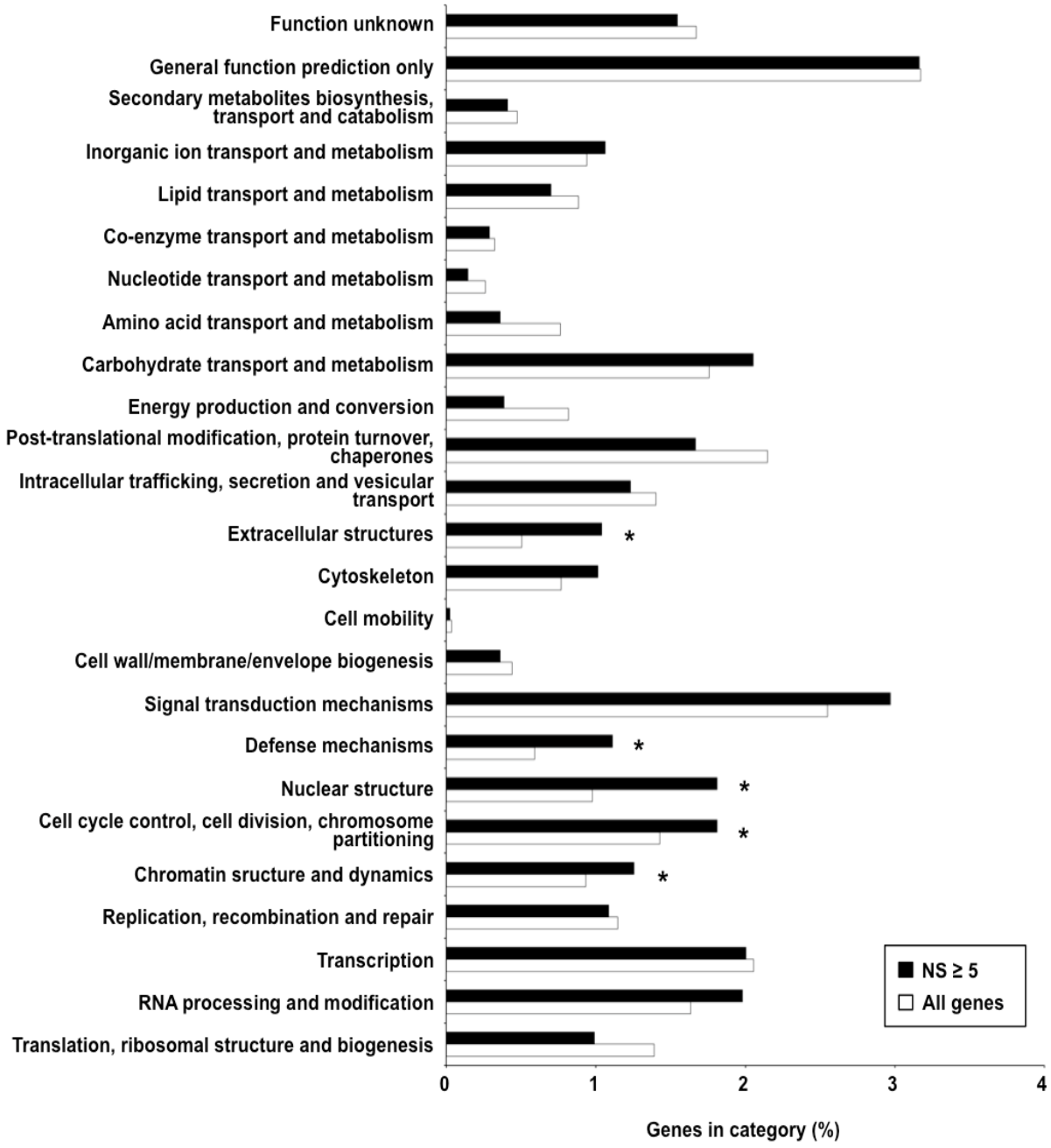


Figure 2: Functional categories over-represented among genes exhibiting five non-synonymous polymorphisms or more. Percentages of genes falling in the different KOG categories among genes exhibiting five non-synonymous polymorphisms or more (NS ≥ 5) relative to the global gene distribution are shown. Black and white bars correspond to selected NS ≥ 5 genes and all genes, respectively. The category "No hits" corresponding to genes with no KOG annotation (~75% in both sets) is not represented on the graph to facilitate visualization of other categories. Significantly over-represented KOG categories are indicated by asterisks (Fisher's exact test, $p < 0.05$).

Protein ID[a]	Protein length	Transcript length	SNP	NS	NS/Kb	Annotation	KOG ID[a]	Go ID[a]
124497	77	231	5	5	21.6	hypothetical secreted protein of 8 kDa	No hit	No hit
124050	151	453	13	9	19.9	hypothetical secreted protein of 17 kDa	No hit	No hit
124361	88	264	5	5	18.9	hypothetical secreted protein of 9 kDa	No hit	No hit
109910	230	690	17	13	18.8	hypothetical secreted protein	No hit	No hit
123541	75	225	6	4	17.8	hypothetical secreted protein of 8 kDa	No hit	No hit
123852	135	405	55	7	17.3	hypothetical secreted protein of 15 kDa	No hit	No hit
104907	117	351	6	6	17.1	hypothetical secreted protein	1245	No hit
123868	139	417	15	7	16.8	hypothetical secreted protein of 15 kDa	No hit	No hit
66458	929	2787	55	45	16.1	hypothetical secreted protein	No hit	No hit
103402	151	453	15	7	15.5	hypothetical secreted protein	No hit	No hit
101262	131	393	18	6	15.3	hypothetical secreted protein	No hit	No hit
124304	200	600	10	9	15.0	hypothetical secreted protein of 22 kDa	No hit	No hit
107425	268	804	28	12	14.9	hypothetical secreted protein	No hit	No hit
124511	67	201	3	3	14.9	hypothetical secreted protein of 7 kDa	No hit	No hit
124264	90	270	5	4	14.8	hypothetical secreted protein of 10 kDa	No hit	9055
107508	720	2160	51	32	14.8	hypothetical secreted protein	No hit	No hit
124351	92	276	7	4	14.5	hypothetical secreted protein of 10 kDa	No hit	No hit
95362	301	903	18	13	14.4	hypothetical secreted protein	No hit	No hit
64885	188	564	23	8	14.2	hypothetical secreted protein of 21 kDa	No hit	No hit
58423	142	426	10	6	14.1	hypothetical secreted protein of 14 kDa	No hit	No hit
124524	71	213	3	3	14.1	hypothetical secreted protein of 8 kDa	No hit	No hit
63656	315	945	22	13	13.8	hypothetical secreted protein	No hit	No hit
70838	97	291	9	4	13.7	hypothetical secreted protein of 10 kDa	No hit	No hit
123559	146	438	10	6	13.7	hypothetical secreted protein of 16 kDa	No hit	No hit
61241	392	1176	39	16	13.6	hypothetical secreted protein	No hit	No hit
68348	247	741	18	10	13.5	hypothetical secreted protein	No hit	No hit
123552	150	450	12	6	13.3	hypothetical secreted protein of 17 kDa	No hit	No hit
124134	125	375	14	5	13.3	hypothetical secreted protein of 14 kDa	No hit	No hit
108793	931	2793	54	37	13.2	hypothetical secreted protein	No hit	No hit
36743	179	537	8	7	13.0	hypothetical secreted protein of 21 kDa	No hit	8237

[a] Protein ID number, Eukaryotic Orthologous Group (KOG) and Gene Ontology (GO) annotations were retrieved from the 98AG31 reference genome at the Joint Genome Institute Mycocosm website (http://genome.jgi.doe.gov/programs/fungi/index.jsf)

Table 6: Top 30 genes encoding secreted proteins accumulating non-synonymous SNPs/Kb.

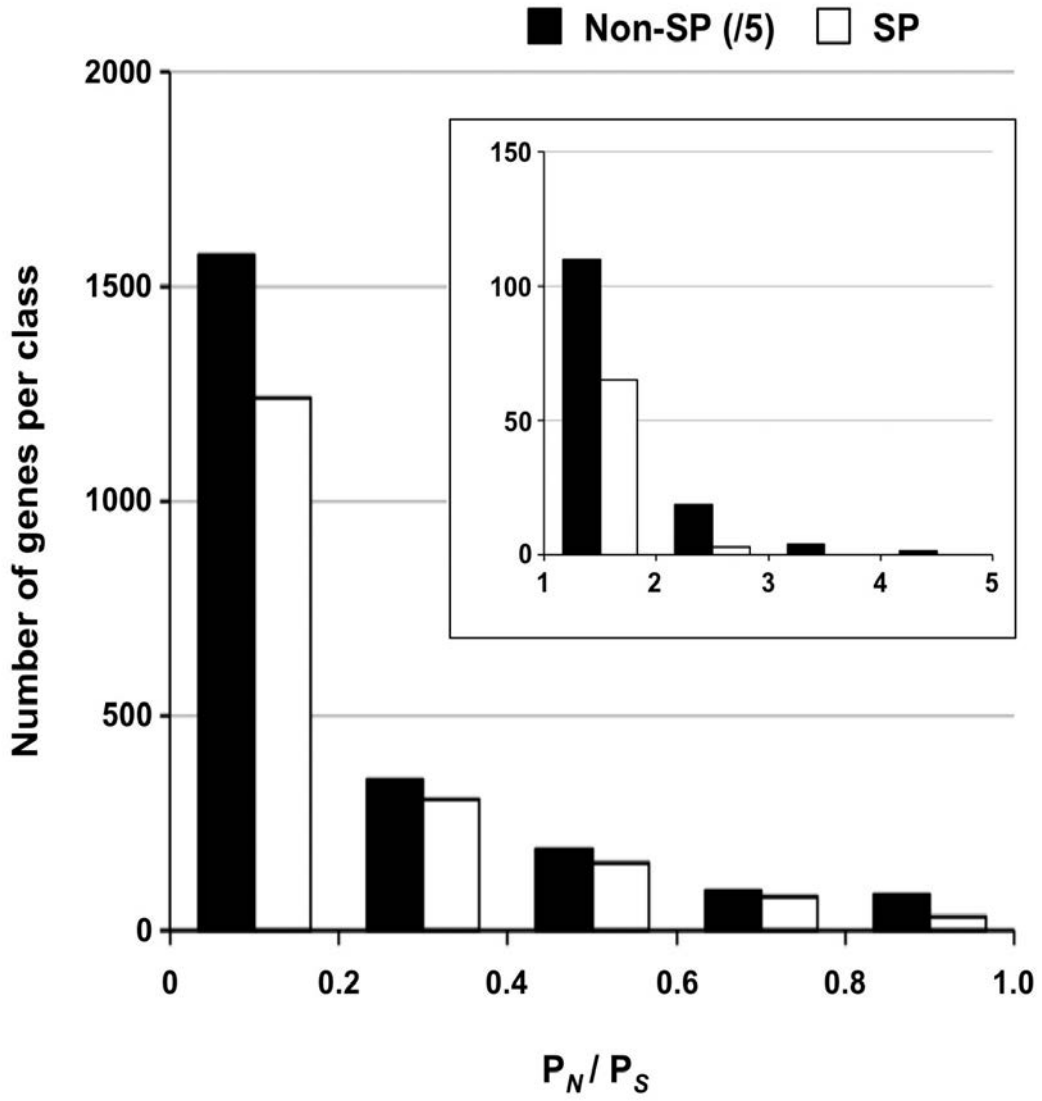


Figure 3: Distribution of P_N/P_S for SP and non-SP genes. Ratios of non-synonymous to synonymous polymorphisms (P_N/P_S) between 0 and 1 are shown for SP genes and non-SP genes. The insert shows distribution of genes with a P_N/P_S >1. Numbers of non-SP genes were divided by 5 for representation. Note the different scale for *y*-axes in figure and insert.

homologs in *Puccinia graminis* f.sp. *tritici* and no conserved domain (ProteinID_124304; Supporting Table 5). The average P_N/P_S observed in SP genes (0.20) was lower than for other genes (0.25). A total of 68 SP genes showed a P_N/P_S >1, whereas 668 had a P_N/P_S>1 in other genes (Supporting Table 6). Among the 30 genes with the highest numbers of non-synonymous SNPs, nine have a P_N/P_S>1 (Table 6). These genes represent particularly interesting candidates that could have evolved under the selection pressure exerted by the interaction with the host plant. No enrichment in KOG functional annotation was detected for the 736 genes presenting a P_N/P_S>1.

In the panel of 15 *M. larici-populina* isolates, only two of the eight virulences described in the poplar rust fungus presented a balanced frequency: virulence 3 with six avirulent isolates and nine virulent isolates and virulence 7 with seven avirulent isolates and eight virulent isolates (Table 1). SP genes presenting conserved non-synonymous SNPs in avirulent isolates and not in virulent isolates (including the reference genome 98AG31 which carries virulences 3 and 7) could be strong candidates, however none of the SP genes presented such a pattern for virulence 3 and 7, suggesting that events other than non-synonymous substitutions in coding sequence may explain the emergence of the virulences 3 and 7. Four SP genes (Protein IDs 89167, 91014, 105154 and 123753) presented non-synonymous SNPs in isolates 98AR1 and 02Y5 which bear the virulence 8, whereas these were absent from the other 13 avirulent isolates, suggesting these genes could be candidate effectors for virulence 8. One SP gene (Protein ID 104703) presented non-synonymous SNPs in isolates 98AR1 and 9683B13 that were absent from the other isolates, indicating that this gene could be a candidate related to virulence 1. One SP gene (Protein ID 108857) presented non-synonymous SNPs in isolates 08EA77, 9683B13 and 09BS12, whereas they were absent from the 12 other isolates, suggesting also that this gene could be a candidate for virulence 8. No correlation was found between mutations in SP genes and other virulences. Similarly, none of the genes interrupted by stop codons correlated with the pathotypes of the 15 isolates.

M. larici-populina SSP genes showing homology to *M. lini Avr* genes *AvrL567*, *AvrP123* and *AvrP4* do not exhibit important accumulation of non-synonymous SNPs (Supporting Table 5). Interestingly, the polymorphic sites identified for the *M. lini* AvrL567 homolog in the poplar rust genome correspond to those that were previously identified by PCR-cloning in a panel of 32 *M. larici-populina* isolates (Hacquard et al., 2012), which included isolate 98AR1, validating the SNPs found in this candidate. Evidence of positive selection were previously recorded for *AvrP4* genes at the intraspecific level in *M. lini* (Barrett et al., 2009) and at the interspecific level in the Melampsoraceae family (Van der Merwe et al., 2009), as well as in a cluster of paralogous genes encoding AvrP4-homologs (multigene family CPG5464; Hacquard et al., 2012). The 13 members of the CPG5464 family in *M. larici-populina* were more closely examined in the 15 isolates (Figure 4). The 13 members of the family were rather conserved and only four had non-synonymous SNPs

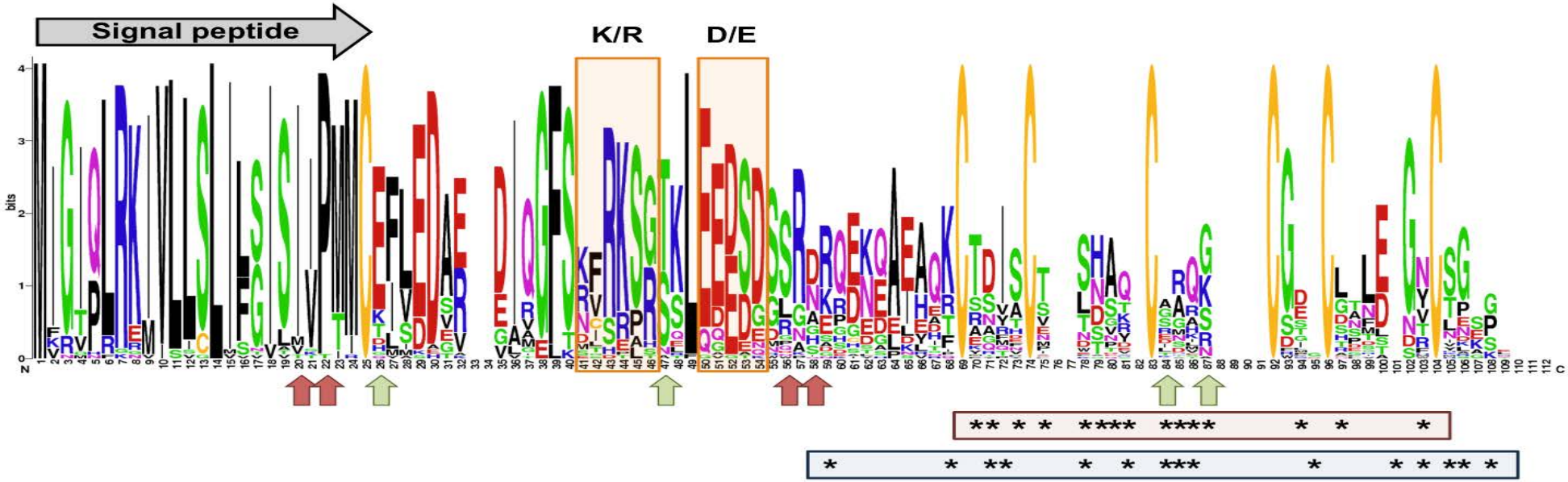


Figure 4: Conservation protein profile of the *M. larici-populina* CPG5464 family and AvrP4 homologs in Melampsoraceae. The profile was designed using WebLogo with 40 sequences corresponding to the 12 members in the CPG5464 family (Hacquard et al. 2012), six variants deduced from the 15 genomes sequenced in this study, 22 AvrP4 homologs sequenced from 9 *Melampsora* spp. (Van der Merwe et al. 2009) and 16 *Melampsora lini* AvrP4 variants. The predicted signal peptide and K/R and D/E rich regions previously shown in Hacquard et al. (2012) are depicted on the profile. Green arrows point to sites under selection in Barrett et al. (2009). Red arrows point to sites of substitution observed in *M. larici-populina* variants. Asterisks in the red box indicate amino acids under positive selection in Van der Merwe *et al.* (2009) and asterisks in the blue box indicate amino acids under positive selection in Hacquard *et al.* (2012).

between isolates (CPG5464_124256, CPG5464_124262, CPG5464_124264, CPG5464_124266). In total, substitutions were noted at four different positions, two within the signal peptide and two after the conserved K/R and E/D regions. None of these substitutions corresponded to positions previously shown under positive selection at the intraspecific or interspecific level (Figure 4). Notably, CPG5464_124564, which includes three different substitution sites in three isolates, presented a P_N/P_S value of 1 and was among the SP genes exhibiting the highest numbers of SNPs/Kb (Table 6, Supporting Table 5). Among the eight homologs of *M. lini AvrM* genes, one showed 15 non-synonymous SNPs (ProteinID_124207; Supporting Table 3). Three *Uromyces fabae* RTP1 homologs have been described in *M. larici-populina* (Hacquard et al., 2012). Only one RTP1 homolog (ProteinID_123932; Supporting Table 3) that consists of a fusion between a *M. lini* HESP-327 homolog and an *U. fabae* RTP1 homolog exhibited an important number of non-synonymous SNPs (7, of which 5 reside in the C-terminal RTP1 region). No substitution occurred at the positions of the four conserved cysteine residues under purifying selection identified by Pretsch et al. (2013).

DISCUSSION

The sequencing of the *M. larici-populina* genome has opened new avenues for the study of effector genes in a model pathosystem composed of a perennial plant and an obligate biotrophic rust fungus (Duplessis et al., 2011a; Hacquard et al., 2011). A set of 1,184 candidate poplar rust effectors were identified on the basis of a combination of typical features of effectors reported in other fungal pathogens, including an initial arbitrary size filter to focus on small proteins of less than 300 amino acids (Hacquard et al., 2012). Because rust fungi effectors such as the *M. lini* AvrM avirulence factor (Ravensdale et al., 2009) can be larger, all predicted secreted proteins were subsequently considered in the search for candidate effectors. Complementary information such as transcript profiling during host infection or the pathogen life cycle can help in reducing the set of genes likely to be *bona fide* effectors (Duplessis et al., 2011b; Hacquard et al., 2013a). Another filter commonly used to identify candidate effectors in plant pathogens is the detection of positive selection in virulence genes, indicative of the evolutionary pressure exerted by the plant-pathogen co-evolution (Alfano et al., 2009; Stergiopoulos and de Wit, 2009). Events such as non-synonymous substitutions, gene gain, gene loss or differential regulation of gene expression can affect avirulence genes and generate new virulences in plant pathogens; comparative genomics using new generation sequencing technologies have uncovered such types of events (Spanu, 2012; Raffaele and Kamoun, 2012). In the present study, we applied Illumina sequencing by synthesis to explore the genetic diversity of *M. larici-populina*, focusing on 15 isolates collected on poplar trees in the wild or in experimental poplar nurseries in the past 21 years in France, and with a wide range of virulence profiles. The main goal here is to provide another level of information about *M. larici-populina*

genes in order to guide selection of pathogenesis-related genes, including effectors, for future functional analyses. The mapping of Illumina reads onto the 98AG31 reference genome helped in the detection of variations such as SNPs and InDels. To date, only a few reports explored genetic diversity at the genome scale in rust fungi using Illumina technology, but they provide ground for comparison within the Pucciniales order (Duplessis et al., 2014b).

Resequencing *M. larici-populina* genomes reveals structural variations

Reads were mapped onto the 98AG31 reference genome with good overall coverage and sequencing depth. Although there was a narrow range in the average coverage by isolate, discrepancies were observed for given scaffolds. Particularly, the small scaffold 484 presented a strikingly high sequencing depth. Two genes encoding an ATP synthase F0 subunit and a NADH dehydrogenase subunit presenting strong similarity with resident genes of the soybean rust *Phakopsora pachyrhizi* mitochondrial genome (Stone et al., 2010) are present on this scaffold. Thus, our analysis identifies a new mitochondrial scaffold that will help in refining the genome assembly. Detailed examination of scaffolds that presented divergent coverage and sequencing depth between isolates revealed on some occasions rather large missing gene-containing regions compared to the reference genome. Although still unresolved, the poplar rust fungus seems to possess a tetrapolar mating system, as for many other basidiomycetes (Duplessis et al., 2011a). In this system, two unlinked loci govern the sexual cycle, and both loci should differ to complete mating (Fraser et al., 2007). Three distinct patterns of conserved missing regions were observed between isolates of unrelated pathotypes collected on different years at different locations (see Table 1 for collection details). Scaffold 90 showed the most striking differences, where missing regions encompass a total of 38 genes, including four pheromone genes that were previously annotated in mating type loci of *M. larici-populina* (Duplessis et al., 2011a). Other mating type loci (i.e. the pheromone/receptor and the homeodomain loci) are also missing in these isolates suggesting that their mating type loci are highly divergent. Despite the quality of the reference genome assembly, the organization of the mating type loci is still not resolved (Duplessis et al., 2011a). This study will provide support to further explore and resolve the organisation and composition of the poplar rust fungus mating loci. Other missing regions unrelated to the mating loci suggest that the poplar rusts posses a great genomic variability. In *M. oryzae*, 1.68 Mb (of a total of 38 Mb) were missing in isolate Ina168 resequenced by 454-pyrosequencing compared to the 70-15 reference genome (Yoshida et al., 2009). This has led to the discovery of many missing SSP genes including known avirulence genes between the two *M. oryzae* isolates (Yoshida et al., 2009). In *M. larici-populina*, none of the missing regions contained large numbers of SP genes (only 12 in total). By performing a wider coverage analysis in the 15 isolates, up to 59 SP genes were found in

low coverage regions, representing possible deletions. However no such deletion correlates with the poplar rust virulences. In *P. striiformis* f. sp. *tritici*, less than 1.3% of the secretome (15 SP genes) was absent between the most divergent sequenced isolates (Cantu et al., 2013), which indicates that the same set of SP genes occurs at the intraspecific level in these rust fungi.

M. larici-populina genomes show remarkable levels of polymorphism

The reference genome 98AG31 was included in the panel of 15 isolates. This genome was previously characterized by Sanger sequencing, which provided an adequate assembly into 462 scaffolds (considering the large size of 101 Mb and a large content in TE, i.e. 45%), however at a rather low sequencing depth of 6.9X (Duplessis et al., 2011a). A total of 88,083 SNPs were previously identified in the reference genome by mapping back Sanger sequencing reads onto the assembled reference genome, with a loose criterion considering a minimum of four reads at a given position (Duplessis et al., 2011a). Illumina sequencing identified a total of 93,189 variants including 86,877 SNPs, of which only 40,001 confirmed SNPs found in the initial assembly. This finding strengthens the support for the use of resequencing at a greater depth to confidently assess SNPs. The total number of SNPs we report is slightly lower than the one found in *P. graminis* f. sp. *tritici* (129,172; Duplessis et al., 2011a). It differs, too, to the numbers reported in *P. striiformis* f. sp. *tritici*, with 81,001 to 108,785 depending on the isolate considered in Zheng et al. (2013) and more than 350,000 with important variations between isolates in Cantu et al. (2013). The large variation in SNPs in these studies could be explained by the wide variation in geographical origin of the isolates and the varying rates of occurrence of sexual reproduction at these sites. Population analyses of the poplar rust fungus with neutral markers indicate that the fungus frequently undergoes sexual recombination resulting in regular gene flow within natural population (Gerard et al., 2006; Barres et al., 2008; Xhaard et al., 2011). Overall, these findings indicate a great genetic diversity in rust fungi that possess a complex life cycle with a sexual reproduction stage achieved on an alternate host (Duplessis et al., 2014b).

Because of the high TE content and the large size of the poplar rust genome, together with putatively large differences between isolates (as previously reported in *P. striiformis* f. sp. *tritici*), we did not expect *de novo* assembly to be optimal for analysis of the 14 isolates sequenced for the first time in this study. Indeed, *de novo* assembly generated large numbers of scaffolds (i.e. > 30,000, data not shown). Instead, Illumina reads from the 14 isolates were directly mapped onto the 98AG31 reference genome for variants detection, similar as in Zheng et al. (2013). In *M. larici-populina*, an average of 148,532 SNPs per isolate were uncovered, which is slightly higher than in *P. striiformis* f. sp. *tritici* according to Zheng et al (2013). The proportions of heterozygous SNPs in the two isolates 08KE26 and 09BS12 (35% and 37%, respectively), might reflect their assignment

to an asexual group as described by the poplar rust population genetic analysis of Xhaard et al. (2011). A much higher proportion of heterozygous SNPs were found between *P. striiformis* f.sp. *tritici* isolates: 82-84% in Zheng et al. (2013) and 87-99% in Cantu et al. (2013). The observed differences between the two studies may reflect differences in the sequencing and analysis process used (Duplessis et al., 2014b), or could be related to a different reproduction regime, as *P. striiformis* f. sp. tritici is mostly asexual which fosters individual heterozygosity (Balloux et al., 2003; Halkett et al., 2005). It would be interesting to compare this with the genetic diversity in rust fungi such as *P. pachyrhizi* or *H. vastatrix* with no known sexual reproduction to date (Rodrigues et al., 1975; Goellner et al., 2009). InDel variants were also inspected and ranged from 4,571 to 8,077 in the 15 *M. larici-populina* isolates, which is slightly larger than in *P. striiformis* f. sp. *tritici* where 1,863 on average were reported (Zheng et al., 2013), but smaller than in the yeast *Saccharomyces* sp. (Liti et al., 2009). A substantial level of polymorphism is noted in *M. larici-populina* at the intraspecific level (~6 SNPs/Kb), which is in close accordance with those reported in the shiitake mushroom *Lentinula edodes* (4.6 SNPs/Kb, Au et al., 2013) or in the wheat stripe rust fungus *P. striiformis* f. sp. *tritici* (Cantu et al., 2013). It is slightly larger than in plant pathogenic ascomycetes such as *Pyrenophora tritici-repentis* (1.9 SNPs/Kb, Manning et al., 2013), *Blumeria graminis* (less than 2 SNPs/Kb; Hacquard et al., 2013b; Wicker et al., 2013) or *Leptosphaeria maculans* (0.5 SNPs/Kb; Zander et al., 2013) but much lower than in the yeast *S. cerevisiae* (59.8 SNPs/Kb; Liti et al., 2009) or in the plant pathogen *Rhizoctonia solani* (~15 SNPs/Kb; Hane et al., 2014). The observed differences in the levels of polymorphism could reflect evolutionary trends related to the lifestyle of these fungi. Rust fungi, exhibit a remarkable level of polymorphism, providing ground for detection of loci that may underlie the co-evolution with their associated hosts and/or their unique life cycle, which is marked by the formation of five spore types and infection of two alternate hosts (Duplessis et al., 2014b).

Patterns of genetic variations in poplar rust genes uncover candidate pathogenesis-related genes

A large part of the variants was identified in coding sequences, similar to *P. striiformis* f. sp. *tritici* (Cantu et al., 2013; Zheng et al., 2013). In total, 89% and 74% of the 16,399 *M. larici-populina* genes were marked by at least one SNP, or one non-synonymous SNP, respectively, in one of the isolates. Such valuable information provides ground for detailed analysis of the functions that may be under selection in the poplar rust genome, particularly those evolving under the pressure of the host plant. P_N/P_S values can be informative to the detection of positive selection and the understanding of how fungi adapt to their environment (Stukenbrock and Bataillon, 2012). We examined the genes showing a $P_N/P_S > 1$ with a particular focus on candidate effectors. Strikingly,

whereas other comparative genomic studies have revealed candidate effector genes under positive selection (Cooke et al., 2012; Wicker et al., 2013), we did not detect any enrichment in SP genes exhibiting a high P_N/P_S compared to all genes in the poplar rust genome. However, 68 SP genes in total showed a $P_N/P_S > 1$ and are priority candidates. Other genes falling in this category may be related to pathogenesis-related functions, but no particular enrichment in functional annotation could be detected. However, the missing regions in *M. larici-populina* isolates contain many genes encoding small proteins (i.e. less than 300 amino acids) with no predicted signal peptide. In the obligate biotroph *B. graminis*, selection analysis carried out between formae speciales identified candidate effectors with no predicted signal peptide that share other common evolutionary features with annotated effectors (Wicker et al., 2013). A total of 262 *M. larici-populina* genes encoding small proteins were found with a $P_N/P_S > 1$ (Supporting Table 6). Such small protein encoding genes are also found among *in planta* highly expressed genes of *M. larici-populina* (Duplessis et al., 2011b). Although no unconventional secretory system is known so far in rust fungi, it would be tempting to consider such proteins in future analysis as possible candidate effectors. We therefore examined the genes presenting a large proportion of non-synonymous substitutions in their sequence and detected enrichment in KOG categories related to nuclear structure and function. Interestingly, genomes of rust fungi contain significantly expanded gene families encoding helicases that may play an important role in DNA repair and maintenance, and nucleic acid and zinc-finger proteins corresponding to putative transcription factors (Duplessis et al. 2011a, Zheng et al. 2013). DNA repair systems can have a dramatic impact on genomic diversity (Seidl and Thomma, 2014) and their possible role in the evolution of the poplar rust genome is still to be determined.

In our study, variations occurring in upstream sequence of genes were also inspected, on the grounds that they may relate to regulation of expression. In total, 16% of the genes had more than 10 SNPs in their 1 Kb upstream region. A detailed analysis of transcriptome-driven analysis of conserved cis-acting regulatory elements in *P. infestans* has revealed motifs underlying specific expression of pathogenesis-related genes (Seidl et al., 2012; Roy et al., 2013a; Roy et al., 2013b). The transcriptome analysis of poplar leaf infection by *M. larici-populina* has shown conserved patterns of coordinated expression of several sets of SSP genes along a time course experiment (Duplessis et al., 2011a). Several other transcriptomic studies have confirmed this trend for SP genes in rust fungi (Fernandez et al., 2012; Cantu et al., 2013; Tremblay et al., 2013; Bruce et al., 2014; Duplessis et al., 2014). A better knowledge of cis-acting regulatory elements in the genome of *M. larici-populina* is needed to further explore the impact of mutations in upstream gene regions. Other molecular mechanisms may control regulation of expression profiles, as recently exemplified in the oilseed rape ascomycete pathogen *L. maculans* (Soyer et al., 2014). Particularly of note, a significant enrichment in genes falling in the chromatin structure and dynamics KOG category was

found in genes accumulating non-synonymous SNPs, and it remains to be explored whether such a control of the chromatin structure could relate to the control of gene expression in rust fungi.

A major goal of the present study was to uncover the presence of polymorphic effectors within a set of predefined candidates that may reflect specific adaptation to the host plant in the classical scheme of the plant-pathogen arms race. A similar approach conducted in *P. striiformis* f. sp. *tritici* identified five polymorphic candidate effectors by comparing two isolates presenting distinct pathotypes (Cantu et al., 2013). Another study identified such possible avirulence genes among secreted protein transcripts showing patterns of non-synonymous mutations between different *Puccinia triticina* isolates (Bruce et al., 2014). In the panel of *M. larici-populina* isolates, virulences 1, 6 and 8 presented correlations with the presence of non-synonymous SNPs in one, one and four genes of virulent isolates compared to avirulent isolates, respectively. Such genes could be candidates underlying virulences 1, 6 and 8. No such correlation was observed for the other virulences carried by the poplar rust isolates, indicating that other events than non-synonymous substitutions in coding sequences may explain their emergence.

Sequence polymorphism has been reported in several avirulence genes of the flax rust *M. lini* (Catanzariti et al., 2006; Dodds et al., 2006; Barett et al., 2009; Van der Merwe et al., 2009; Ravensdale et al., 2011). Homologs of flax rust avirulence genes retrieved in the *M. larici-populina* genome did not exhibit high P_N/P_S or excess of non-synonymous substitutions in the 15 isolates, except in a very few cases. Interestingly, non-synonymous substitutions observed in the CPG5464 family homologous to *M. lini* AvrP4 did not match sites previously shown under selection in *M. lini* at the intraspecific level (Barrett et al., 2009), in Melampsoraceae at the interspecific level (Van der Merwe et al., 2009) or between members of the paralogous gene cluster CPG5464 of *M. larici-populina* (Hacquard et al., 2012). Members of this gene family are rather conserved within the Melampsoraceae, suggesting that AvrP4/CPG5464 could play an important role as an effector during the interaction with the relative host plants. A high diversity is observed at both the intraspecific and interspecific level highlighting the probable interplay with the different host plants, but to date, such an interaction in a gene-for-gene manner has only been demonstrated for the flax rust fungus (Ravensdale et al., 2011). At least one *M. larici-populina* homolog of the *M. lini AvrM* gene shows a high level of polymorphic sites (e.g. in isolate 98AR1, 30 SNPs of which 15 are non-synonymous), similar to those reported in *M. lini* (Catanzariti et al., 2006; Ravensdale et al., 2011). Some of these mutations are particularly important for the direct interaction with the corresponding *M* resistance gene in flax (Catanzariti et al., 2010; Ve et al., 2013). It will be particularly interesting to further study the potential role of AvrM homologs in the poplar-poplar rust fungus interaction.

Future steps in poplar rust genomics

Genomics is a powerful approach to identify pathogenesis-related candidates, as the present study illustrates. From the perspective of population biology, it is well known that structure and demography can affect all loci equally. To identify loci under selection, a population genomics approach is required to take into account demographic history. A population genomics study is ongoing in collaboration with the JGI to identify loci related to virulence 7. As large portions of the genome were missing in different *M. larici-populina* isolates, it might be required to study presence/absence at a larger scale using *de novo* assembled genomes. Many mechanisms can underlie genome evolution (Raffaele and Kamoun, 2012; Seidl and Thomma, 2014) and a better knowledge of the structural rearrangements occurring in the poplar rust genome will help to determine their impact on virulence evolution. In this regard, we have initiated the genome sequencing of an avirulent 7 isolate by combining paired-end and mate-pair Illumina sequencing to compare with the virulent 7 reference genome. Together, these genomic analyses will foster functional studies by pinpointing numerous sites of sequence variation, i.e. positions that may have important implications at the structural level for the function of effectors.

Author contributions

Sébastien Duplessis and Pascal Frey designed research; Antoine Persoons, Sébastien Duplessis, Christine Delaruelle, and Pascal Frey performed research; Antoine Persoons, Sébastien Duplessis, Emmanuelle Morin, Stéphane De Mita and Thibaut Payen analyzed data; Antoine Persoons and Sébastien Duplessis drafted the manuscript and, Antoine Persoons, Sébastien Duplessis, Pascal Frey, Fabien Halkett, Thibaut Payen and Stéphane De Mita wrote the paper.

Acknowledgments

We warmly thank Katherine Hayden for comments on the manuscript. We would like to acknowledge the help of Bénédicte Fabre (INRA Nancy) in the production of poplar plants in greenhouses and of *M. larici-populina* urediniospores. We also thank our colleagues Claude Murat and Francis Martin at INRA Nancy for fruitful discussions during the course of the study. This work was supported by the French National Research Agency through the Laboratory of Excellence ARBRE (ANR-12-LABXARBRE-01), the Young Scientist Grant POPRUST to Sébastien Duplessis (ANR-2010-JCJC-1709-01) and the GANDALF project (ANR-12-ADAP0009) and by the Région Lorraine (Researcher Award to Sébastien Duplessis). Antoine Persoons is supported by a Doctoral Scholarship from the Institut National de la Recherche Agronomique and the Region Lorraine. We thank the Joint Genome Institute for the access to the *M. larici-populina* genome sequence. JGI sequencing is supported by the Office of Science of the United States Department of Energy under

contract number DE-AC02-05CH11231.

REFERENCES

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25:3389-3402.

Alfano, J.R. (2009). Roadmap for future research on plant pathogen effectors. Mol Plant Pathol. 10:805-13. doi: 10.1111/j.1364-3703.2009.00588.x.

Au, C.H., Cheung, M.K., Wong, M.C., Chu, A.K., Law, P.T., Kwan, H.S. (2013). Rapid genotyping by low-coverage resequencing to construct genetic linkage maps of fungi: a case study in *Lentinula edodes*. BMC Res Notes. 6:307. doi: 10.1186/1756-0500-6-307.

Balloux, F., Lehmann, L., De Meeus, T., (2003). The population genetics of clonal and partially clonal diploids. Genetics 164, 1635-1644.

Barrès B, Halkett F, Dutech C, Andrieux A, Pinon J, Frey P. (2008). Genetic structure of the poplar rust fungus *Melampsora larici-populina*: evidence for isolation by distance in Europe and recent founder effects overseas. *Infect Genet Evol*. 8(5):577-587 doi: 10.1016/j.meegid.2008.04.005

Barrett LG, Thrall PH, Dodds PN, van der Merwe M, Linde CC, Lawrence GJ, Burdon JJ. (2009). Diversity and evolution of effector loci in natural populations of the plant pathogen *Melampsora lini*. *Mol Biol Evol*. 26(11):2499-2513 doi:10.1093/molbev/msp166

Bruce M, Neugebauer KA, Joly DL, Migeon P, Cuomo CA, Wang S, Akhunov E, Bakkeren G, Kolmer JA, Fellers JP. (2014). Using transcription of six *Puccinia triticina* races to identify the effective secretome during infection of wheat. *Front Plant Sci*. 4:520 doi: 10.3389/fpls.2013.00520

Cantu D, Govindarajulu M, Kozik A, Wang M, Chen X, Kojima KK, Jurka J, Michelmore RW, Dubcovsky J. (2011). Next generation sequencing provides rapid access to the genome of *Puccinia striiformis f. sp. tritici*, the causal agent of wheat stripe rust. *PLoS One*. 6(8):e24230 doi: 10.1371/journal.pone.0024230

Cantu D, Segovia V, MacLean D, Bayles R, Chen X, Kamoun S, Dubcovsky J, Saunders DG, Uauy C. (2013). Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis f. sp. tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. *BMC Genomics*. 14:270 doi: 10.1186/1471-2164-14-270

Catanzariti, A-M., Dodds, P.N., Lawrence, G.J., Ayliffe, M.A., Ellis, J.G. (2006) Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. Plant Cell. 18:243-56.

Catanzariti, A-M., Dodds, P.N., Ve, T., Kobe, B., Ellis, J.G., Staskawicz, B.J. (2010). The AvrM effector from flax rust has a structured C-terminal domain and interacts directly with the M resistance protein. Mol Plant Microbe Interact. 23:49-57. doi: 10.1094/MPMI-23-1-0049.

Cooke, D.E., Cano, L.M., Raffaele, S., Bain, R.A., Cooke, L.R., Etherington, G.J., Deahl, K.L., Farrer, R.A., Gilroy, E.M., Goss, E.M., Grünwald, N.J., Hein, I., MacLean, D., McNicol, J.W., Randall, E., Oliva, R.F., Pel, M.A., Shaw, D.S., Squires, J.N., Taylor, M.C., Vleeshouwers, V.G., Birch, P.R., Lees, A.K., Kamoun, S. (2012). Genome analyses of an aggressive and invasive lineage of the Irish potato famine pathogen. PLoS Pathog. 8:e1002940. doi: 10.1371/journal.ppat.1002940.

Crooks GE, Hon G, Chandonia JM, Brenner SE.(2004). WebLogo: a sequence logo generator. *Genome Res*. 14(6):1188-90.

De Mita S, Siol M. (2012) EggLib: processing, analysis and simulation tools for population genetics and genomics. *BMC Genet.* 11:13-27. doi: 10.1186/1471-2156-13-27

Dodds, P.N., Lawrence, G.J., Catanzariti, A-M., Ayliffe, M.A., Ellis, J.G. (2004) The *Melampsora lini* AvrL567 avirulence genes are expressed in haustoria and their products are recognized inside plant cells. Plant Cell. 16:755-68.

Dodds, P.N., Lawrence, G.J., Catanzariti, A-M., Teh, T., Wang, C.I., Ayliffe, M.A., Kobe, B., Ellis, J.G. (2006). Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. Proc Natl Acad Sci U S A. 103:8888-93.

Dodds PN, Rathjen JP. (2010). Plant immunity: towards an integrated view of plant-pathogen interactions. *Nat Rev Genet.* 11(8):539-548 doi:10.1038/nrg2812

Duplessis S, Cuomo CA, Lin YC, Aerts A, Tisserant E, Veneault-Fourrey C, Joly DL, Hacquard S, Amselem J, Cantarel BL, Chiu R, Coutinho PM, Feau N, Field M, Frey P, Gelhaye E, Goldberg J, Grabherr MG, Kodira CD, Kohler A, Kües U, Lindquist EA, Lucas SM, Mago R, Mauceli E, Morin E, Murat C, Pangilinan JL, Park R, Pearson M, Quesneville H, Rouhier N, Sakthikumar S, Salamov AA, Schmutz J, Selles B, Shapiro H, Tanguay P, Tuskan GA, Henrissat B, Van de Peer Y, Rouzé P, Ellis JG, Dodds PN, Schein JE, Zhong S, Hamelin RC, Grigoriev IV, Szabo LJ, Martin F. (2011a). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. Proc Natl Acad Sci USA. 108:9166-9171 doi: *10.1073/pnas.1019315108*

Duplessis S, Hacquard S, Delaruelle C, Tisserant E, Frey P, Martin F, Kohler A. (2011b). *Melampsora larici-populina* transcript profiling during germination and timecourse infection of poplar leaves reveals dynamic expression patterns associated with virulence and biotrophy. *Mol Plant Microbe Interact.* 24(7):808-818 doi: 10.1094/MPMI-01-11-0006

Duplessis S, Spanu PD, Schirawski J. (2014a). Biotrophic fungi (powdery mildews, Rusts and Smuts). In Martin F. *Ecological genomics of the fungi. Plant-interacting fungi section*. Wiley-Blackwell. *149-168.*

Duplessis, S., Bakkeren, G., Hamelin, R. (2014b). Advancing knowledge on biology of rust fungi through genomics. *Advances in Botanical Research* 70:173-209.

Fernandez D, Tisserant E, Talhinhas P, Azinheira H, Vieira A, Petitot AS, Loureiro A, Poulain J, Da Silva C, Silva Mdo C, Duplessis S.(2012). 454-pyrosequencing of Coffea arabica leaves infected by the rust fungus *Hemileia vastatrix* reveals in planta-expressed pathogen-secreted proteins and plant functions in a late compatible plant-rust interaction. *Mol Plant Pathol*. 13(1):17-37. doi: 10.1111/j.1364-3703.2011.00723

Flor, H. H. (1971). Current Status of the Gene-For-Gene Concept. *Annu. Rev. Phytopathol.* 9:275–296.

Fraser JA, Hsueh YP, Findley KM and Heiman J (2007) Evolution of the mating type locus: the basidiomycetes. In “Sex in fungi: molecular determination and evolutionary implications”eds Heitman J, Kronstad JW, Taylor JW and Casselton LA, ASM Press, Washington DC, 19-34

Gérard PR, Husson C, Pinon J, Frey P. (2006). Comparison of genetic and virulence diversity of *Melampsora larici-populina* populations on wild and cultivated poplar and influence of the alternate host. *Phytopathology.* 96:1027-1036 doi: 10.1094/PHYTO-96-1027

Giraldo MC, Valent B. (2013). Filamentous plant pathogen effectors in action. Nat Rev Microbiol. 11(11):800-814 doi: 10.1038/nrmicro3119

Goellner, K., Loehrer, M., Langenbach, C., Conrath, U. W. E., Koch, E., and Schaffrath, U. (2009). *Phakopsora pachyrhizi*, the causal agent of Asian soybean rust. Mol. Plant Pathol. 11, 169–177

Hacquard S, Petre B, Frey P, Hecker A, Rouhier N, Duplessis S. (2011). The poplar-poplar rust interaction: insights from genomics and transcriptomics. *JPathog.* 2011:716041 doi: 10.4061/2011/716041

Hacquard S, Joly D.L, Lin Y.C, Tisserant E, Feau N, Delaruelle C, Legué V, Kohler A, Tanguay T,

Petre B, Frey P, Van de Peer Y, Rouzé P, Martin F, Hamelin R.C, Duplessis S. (2012). A comprehensive analysis of genes encoding small secreted proteins identifies candidate effectors in *Melampsora larici-populina* (Poplar Leaf Rust). *Mol Plant Microbe Interact.* 25:279-293 doi: 10.1094/MPMI-09-11-0238

Hacquard, S., Delaruelle, C., Frey, P., Tisserant, E., Kohler, A., Duplessis, S. (2013). Transcriptome analysis of poplar rust telia reveals overwintering adaptation and tightly coordinated karyogamy and meiosis processes. Front Plant Sci. 4:456. doi: 10.3389/fpls.2013.00456.

Hacquard, S., Kracher, B., Maekawa, T., Vernaldi, S., Schulze-Lefert, P., Ver Loren van Themaat, E. (2013b). Mosaic genome structure of the barley powdery mildew pathogen and conservation of transcriptional programs in divergent hosts. Proc Natl Acad Sci U S A 110:E2219-28. doi: 10.1073/pnas.1306807110.

Halkett, F., Simon, J-C., Balloux, F. (2005). Tackling the population genetics of clonal and partially clonal organisms. Trends Ecol Evol 20:194-201. DOI: 10.1016/j.tree.2005.01.001

Hane, J.K., Anderson, J.P., Williams, A.H., Sperschneider, J., Singh, K.B. (2014). Genome sequencing and comparative genomics of the broad host-range pathogen *Rhizoctonia solani* AG8. PLoS Genet. 10:e1004281. doi: 10.1371/journal.pgen.1004281.

Joly DL, Feau N, Tanguay P, Hamelin RC. (2010). Comparative analysis of secreted protein evolution using expressed sequence tags from four poplar leaf rusts (*Melampsora* spp.). *BMC Genomics*. 11:422 doi: 10.1186/1471-2164-11-422

Kemen E, Kemen AC, Rafiqi M, Hempel U, Mendgen K, Hahn M, Voegele RT. (2005). Identification of a protein from rust fungi transferred from haustoria into infected plant cells. *Mol Plant Microbe Interact*. 18(11):1130-1139.

Kemen E, Kemen A, Ehlers A, Voegele R, Mendgen K. (2013). A novel structural effector from rust fungi is capable of fibril formation. *Plant J*. 75(5):767-780 doi: 10.1111/tpj.12237

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, Tsai IJ, Bergman CM, Bensasson D, O'Kelly MJ, van Oudenaarden A, Barton DB, Bailes E, Nguyen AN, Jones M, Quail MA, Goodhead I, Sims S, Smith F, Blomberg A, Durbin R, Louis EJ. (2009). Population genomics of domestic and wild yeasts. *Nature*. 458(7236):337-341. doi: 10.1038/nature07743

Lowe RG, Howlett BJ. (2012). Indifferent, affectionate, or deceitful: lifestyles and secretomes of

fungi. *PLoS Pathog*. 8(3):e1002515 doi: 10.1371/journal.ppat.1002515

Manning, V.A., Pandelova, I., Dhillon, B., Wilhelm, L.J., Goodwin, S.B., Berlin, A.M., Figueroa, M., Freitag, M., Hane, J.K., Henrissat, B., Holman, W.H., Kodira, C.D., Martin, J., Oliver, R.P., Robbertse, B., Schackwitz, W., Schwartz, D.C., Spatafora, J.W., Turgeon, B.G., Yandava, C., Young, S., Zhou, S., Zeng, Q., Grigoriev, I.V., Ma, L.J., Ciuffetti, L.M. (2013). Comparative genomics of a plant-pathogenic fungus, *Pyrenophora tritici-repentis*, reveals transduplication and the impact of repeat elements on pathogenicity and population divergence. G3. 3:41-63. doi: 10.1534/g3.112.004044.

Nemri A, Saunders DGO, Anderson C, Upadhyaya NM, Win J, Lawrence GJ, Jones DA, Kamoun S, Ellis JG, Dodds PN. (2014). The genome sequence and effector complement of the flax rust pathogen *Melampsora lini*. *Front. Plant Sci*. 5:98 doi: 10.3389/fpls.2014.00098

Pinon J, Frey P. (2005). Interactions between poplar clones and *Melampsora* populations and their implications for breeding for durable resistance. In Pei M. H., McCracken A. R., eds. Rust diseases of Willow and Poplar. *CAB International*, Wallingford, UK139-154.

Pretsch K, Kemen A, Kemen E, Geiger M, Mendgen K, Voegele R. (2013). The rust transferred proteins-a new family of effector proteins exhibiting protease inhibitor function. *Mol Plant Pathol*. 14(1):96-107 doi: 10.1111/j.1364-3703.2012.00832

Raffaele S, Kamoun S. (2012). Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat Rev Microbiol*. 10(6):417-30 doi: 10.1038/nrmicro2790

Ravensdale M, Nemri A, Thrall PH, Ellis JG, Dodds PN. (2011). Co-evolutionary interactions between host resistance and pathogen effector genes in flax rust disease. *Mol Plant Pathol*. 12(1):93-102 doi: 10.1111/j.1364-3703.2010.00657

Rodrigues C.J., Jr, Bettencourt, A.J. and Rijo, L. (1975) Races of the patho- gen and resistance to coffee rust. Annu. Rev. Phytopathol. 13, 49–70.

Roy, S., Kagda, M., Judelson, H.S. (2013a). Genome-wide prediction and functional validation of promoter motifs regulating gene expression in spore and infection stages of *Phytophthora infestans*. PLoS Pathog. 9:e1003182. doi: 10.1371/journal.ppat.1003182.

Roy, S., Poidevin, L., Jiang, T., Judelson, H.S. (2013b). Novel core promoter elements in the oomycete pathogen Phytophthora infestans and their influence on expression detected by genome-wide analysis. BMC Genomics. 14:106. doi: 10.1186/1471-2164-14-106.

Saunders, D.G.O., Win, J., Cano, L.M., Szabo, L.J., Kamoun, S., Raffaele, S. (2012). Using hierarchical clustering of secreted protein families to classify and rank candidate effectors of rust fungi. *PLoS ONE* 7(1): e29847 doi: 10.1371/journal.pone.0029847

Seidl, M.F., Wang, R-P., Van den Ackerveken, G., Govers, F., Snel., B. (2012). Bioinformatic inference of specific and general transcription factor binding sites in the plant pathogen *Phytophthora infestans*. *PLoS ONE* 7(12): e51295. doi: 10.1371/journal.pone.0051295

Seidl, M.F., Thomma, B.P.H.J. (2014). Sex or no sex: Evolutionary adaptation occurs regardless. BioEssays DOI: 10.1002/bies.201300155

Soyer JL, El Ghalid M, Glaser N, Ollivier B, Linglin J, Grandaubert J, Balesdent MH, Connolly LR, Freitag M, Rouxel T, Fudal I. (2014). Epigenetic control of effector gene expression in the plant pathogenic fungus *Leptosphaeria maculans*. *PLoS Genet.* 10(3):e1004227. doi: 10.1371/journal.pgen.1004227

Spanu, P.D. (2012). The genomics of obligate (and nonobligate) biotrophs. Annu Rev Phytopathol. 50:91-109. doi: 10.1146/annurev-phyto-081211-173024.

Steenackers J, Steenackers M, Steenackers V, Stevens M. (1996) Poplar diseases, consequences on growth and wood quality. Biomass and Bioenergy. 10(5–6):267–274.

Stergiopoulos I, de Wit PJ. (2009). Fungal effector proteins. *Annual Review of Phytopathology* 47: 233-263 doi: 10.1146/annurev.phyto.112408.132637

Stergiopoulos I, Cordovez V, Okmen B, Beenen HG, Kema GH, de Wit PJ. (2013). Positive selection and intragenic recombination contribute to high allelic diversity in effector genes of *Mycosphaerella fijiensis*, causal agent of the black leaf streak disease of banana. *Mol Plant Patho*l. 15(5):447-60 doi: 10.1111/mpp.12104

Stone CL, Buitrago ML, Boore JL, Frederick RD. (2010). Analysis of the complete mitochondrial genome sequences of the soybean rust pathogens *phakopsora pachyrhizi* and *p. meibomiae*. *Mycologia*.102(4):887-97.

Stukenbrock EH, McDonald BA. (2009). Population genetics of fungal and oomycete effectors involved in gene-for-gene interactions. *Mol Plant Microbe Interact*. 22(4):371-80 doi: 10.1094/MPMI-22-4-0371

Stukenbrock EH, Bataillon T. (2012). A population genomics perspective on the emergence and adaptation of new plant pathogens in agro-ecosystems. *PLoS Pathog*. 8(9):e1002893 doi:

10.1371/journal.ppat.1002893

Stukenbrock EH. (2013). Evolution, selection and isolation: a genomic view of speciation in fungal plant pathogens. *New Phytol*. 199(4):895-907 doi: 10.1111/nph.12374

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*. (11):4-41.

Thompson JD, Gibson TJ, Higgins DG. (2002). Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics*. Chapter 2:Unit 2.3. doi: 10.1002/0471250953. bi0203s00

Tremblay A, Hosseini P, Li S, Alkharouf NW, Matthews BF. (2013). Analysis of *Phakopsora pachyrhizi* transcript abundance in critical pathways at four time-points during infection of a susceptible soybean cultivar using deep sequencing. *BMC Genomics*. 11:14-614. doi: 10.1186/1471-2164-14-614

Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R.R., Bhalerao, R.P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.L., Cooper, D., Coutinho, P.M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Déjardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I.,Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjärvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leplé, J-C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D.R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouzé, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y., Rokhsar, D. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science. 313:1596-604.

Tyler BM and Rouxel T. (2012) Effectors of fungi and oomycetes: their virulence and avirulence functions and translocation from pathogen to host cells, in *Molecular Plant Immunity* (ed G. Sessa), Wiley-Blackwell, Oxford, UK. doi: 10.1002/9781118481431.ch7

Upadhyaya NM, Mago R, Staskawicz BJ, Ayliffe MA, Ellis JG, Dodds PN. (2014). A bacterial type III secretion assay for delivery of fungal effector proteins into wheat. *Mol Plant Microbe Interact*. 27(3):255-264 doi: 10.1094/MPMI-07-13-0187-FI

Van der Merwe MM, Kinnear MW, Barrett LG, Dodds PN, Ericson L, Thrall PH, Burdon JJ. (2009). Positive selection in AvrP4 avirulence gene homologues across the genus *Melampsora*. *Proc Biol Sci*. 276(1669):2913-2922 doi: 10.1098/rspb.2009.0328

Ve, T., Williams, S.J., Catanzariti, A-M., Rafiqi, M., Rahman, M., Ellis, J.G., Hardham, A.R., Jones, D.A., Anderson, P.A., Dodds, P.N., Kobe, B. (2013). Structures of the flax-rust effector AvrM reveal insights into the molecular basis of plant-cell entry and effector-triggered immunity. Proc Natl Acad Sci U S A. 110:17594-9. doi: 10.1073/pnas.1307614110.

Wicker, T., Oberhaensli, S., Parlange, F., Buchmann, J.P., Shatalina, M., Roffler, S., Ben-David, R., Doležel, J., Šimková, H., Schulze-Lefert, P., Spanu, P.D., Bruggmann, R., Amselem, J., Quesneville, H., Ver Loren van Themaat, E., Paape, T., Shimizu, K.K., Keller, B. (2013). The wheat powdery mildew genome shows the unique evolution of an obligate biotroph. Nat Genet. 45:1092-6. doi: 10.1038/ng.2704.

Win J, Chaparro-Garcia A, Belhaj K, Saunders DG, Yoshida K, Dong S, Schornack S, Zipfel C, Robatzek S, Hogenhout SA, Kamoun S. (2012). Effector biology of plant-associated organisms: concepts and perspectives. *Cold Spring Harb Symp Quant Biol*. 77:235-247 doi: 10.1101/sqb.2012.77.015933

Xhaard C, Fabre B, Andrieux A, Gladieux P, Barrès B, Frey P, Halkett F. (2011). The genetic structure of the plant pathogenic fungus *Melampsora larici-populina* on its wild host is extensively impacted by host domestication. *Mol Ecol.* 20:2739-2755 doi: 10.1111/j.1365-294X.2011.05138

Yoshida K, Saitoh H, Fujisawa S, Kanzaki H, Matsumura H, Yoshida K, Tosa Y, Chuma I, Takano Y, Win J, Kamoun S, Terauchi R. (2009). Association genetics reveals three novel avirulence genes from the rice blast fungal pathogen Magnaporthe oryzae. *Plant Cell.* 21(5):1573-1591. doi: 10.1105/tpc.109.066324

Zander, M., Patel, D.A., Van de Wouw, A., Lai, K., Lorenc, M.T., Campbell, E., Hayward, A., Edwards, D., Raman, H., Batley, J. (2013). Identifying genetic diversity of avirulence genes in *Leptosphaeria maculans* using whole genome sequencing. Funct Integr Genomics. 13:295-308. doi: 10.1007/s10142-013-0324-5.

Zheng W, Huang L, Huang J, Wang X, Chen X, Zhao J, Guo J, Zhuang H, Qiu C, Liu J, Liu H, Huang

X, Pei G, Zhan G, Tang C, Cheng Y, Liu M, Zhang J, Zhao Z, Zhang S, Han Q, Han D, Zhang H, Zhao J, Gao X, Wang J, Ni P, Dong W, Yang L, Yang H, Xu JR, Zhang G, Kang Z. (2013) High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus. *Nat Commun*. 4:2673 doi: 10.1038/ncomms3673

Supporting Tables

Les données supplémentaires sont accessibles à l'adresse suivante en téléchargement (un dossier excel) : http://journal.frontiersin.org/article/10.3389/fpls.2014.00450/full

Supporting Table 1. Detailed analysis of 151 *M. larici-populina* scaffolds presenting discrepancies of average sequencing depth between scaffolds and/or between the 15 isolates surveyed in the study. The resolution columns indicate whether the discrepancies could relate to the presence of transposable elements (TE), sequencing depth difference, missing regions in the scaffold in one or several of the isolates and missing genes (including SP genes) present in these regions.

Supporting Table 2. Coverage analysis identifying high and low coverage regions in the genomes of *M. larici-populina* isolates compared to the 98AG31 reference genome at JGI. The numbers of SP genes present in the low coverage regions are indicated.

Supporting Table 3. Detailed analysis of variants identified in all genes in the 15 *M. larici-populina* isolates when compared to the 98AG31 reference genome at JGI. Numbers of variants such as SNPs, non-synonymous SNPs (NS), SNPs in coding DNA sequence (SNP_CDS), MNVs and insertions/deletions are provided, either as total numbers considering all isolates or in each isolate. Details of SNPs, MNVs, insertions and deletions identified in the 1 Kb upstream regions are also provided. Other parameters measured in the study like P_N/P_S or numbers of variants per Kb are detailed.

Supporting Table 4. Top 30 genes accumulating Single Nucleotide Polymorphism (SNP) in their 1 Kb upstream region. Gene Ontology (GO) and Eukaryotic Orthologous Group (KOG) IDs of corresponding proteins were retrieved on the 98AG31 reference genome JGI website.

Supporting Table 5. Detailed analysis of variants identified in all SP genes in the 15 *M. larici-populina* isolates when compared to the 98AG31 reference genome at JGI. Numbers of variants such as SNPs, non-synonymous SNPs (NS), SNPs in coding DNA sequence (SNP_CDS), MNVs and insertions/deletions are provided, either as total numbers considering all isolates or in each isolate. Details of SNPs, MNVs, insertions and deletions identified in the 1 Kb upstream regions are

also provided. Other parameters measured in the study like P_N/P_S or numbers of variants per Kb are detailed.

Supporting Table 6. Detailed numbers of SNPs and non-synonymous SNPs (NS) for 736 *M. larici-populina* genes presenting a $P_N/P_S > 1$.

Chapitre 3

Impact du contournement de la résistance 7 sur les populations de *Melampsora larici-populina*

Chapitre 3 - Impact du contournement de la résistance 7 sur les populations de *Melampsora larici-populina*

1 Introduction

Une étude antérieure a permis de montrer que l'évènement de contournement de la résistance 7 avait largement façonné la structure génétique des populations de ce pathogène, avec une forte différenciation entre des individus sauvages (incapables de contourner la résistance 7 et originaires de la zone de distribution originelle du Peuplier noir) et des individus virulents 7 (Xhaard et al., 2011).

L'objectif de ce chapitre est de décrire l'effet du contournement de la résistance 7 sur l'histoire démographique de *M. larici-populina* à l'aide d'un jeu de données temporel. Pour cela, 594 individus français et belges échantillonnés entre 1992 et 2012 on été génotypés, en incluant des individus issus de l'analyse de Xhaard et al., (2011). A l'aide du logiciel TESS (Durand et al., 2009), trois groupes génétiques ont été identifiés au sein de nos isolats. Parmi eux, nous avons retrouvé les groupes des sauvages et des virulents 7 précédemment décrits (Xhaard et al., 2011). Le groupe qualifié de virulent 7 (post-contournement) n'a pas été échantillonné avant 1994 et correspond à des individus présentant majoritairement la virulence 7. Il est présent dans les zones géographiques de plantation du cultivar Beaupré (porteur de la résistance 7). Le groupe sauvage regroupe les individus échantillonnés dans la zone originelle de sympatrie peuplier-mélèze et ne présentant pas la virulence 7. De plus, un nouveau groupe génétique a été identifié, que nous qualifierons ici d'avirulent 7 (pré-contournement) dont on ne retrouve plus de représentant après 1997 et qui correspond à des individus majoritairement avirulents 7 et échantillonnés dans les zones de plantations de Beaupré.

En utilisant l'assignation des individus aux différents groupes génétiques pour définir des populations, nous avons pu examiner les caractéristiques génétiques de ces populations. L'évolution de la richesse allélique et de la diversité génétique au cours du temps a été mesurée. Le groupe sauvage présente la plus grande diversité génétique, le groupe des avirulents 7 la plus faible et le groupe des virulents 7 présente des valeurs différentes en fonction de l'année d'échantillonnage, partant d'une valeur faible en 1994 et qui augmente jusqu'à se stabiliser vers 2006. Ce résultat pourrait être le signe d'un goulot d'étranglement suivi d'une expansion démographique chez les virulents 7 ce qui a été confirmé par l'utilisation du logiciel Bottleneck (Cornuet and Luikart, 1996).

Ce chapitre se présente sous le forme d'un article en préparation.

2. Article n°2 : Article n°2 : **Population replacement following a major selection event in the plant pathogen *Melampsora larici-populina***

Antoine Persoons[1,2], Bénédicte Fabre[1,2], Pascal Frey[1,2], Stéphane De Mita[1,2] and Fabien Halkett[1,2]

[1]INRA, Unité Mixte de Recherche 1136 INRA/Université de Lorraine, Interactions Arbres/Microorganismes, 54280 Champenoux, France
[2]Université de Lorraine, Unité Mixte de Recherche 1136 INRA/Université de Lorraine, 54506 Vandoeuvre-lès-Nancy Cedex, France

Corresponding author:
Dr Fabien Halkett
INRA, Unité Mixte de Recherche 1136 INRA/Université de lorraine,
Interactions Arbres/Microorganismes, 54280 Champenoux, France
e-mail, halkett@nancy.inra.fr

ABSTRACT

Melampsora larici-populina is a fungal pathogen responsible for foliar rust disease on poplar trees causing important damage worldwide and particularly in plantations of northern Europe. Breeding and massive deployment, in France and Europe, of poplar clones carrying qualitative resistance have greatly impacted the population genetic structure of *M. larici-populina* (Xhaard et al., 2011). All resistances deployed to date have been overcome. A major selection event occurred in 1994 with the breakdown of resistance R7 carried by some poplar cultivars (such as Beaupré) massively planted in France since the 80's. The corresponding virulence rapidly spread in *M. larici-populina* populations, and nearly reached fixation in northern France, even on susceptible hosts.

Using an extensive collection of nearly 600 *M. larici-populina* individuals sampled in France and Belgium from 1992 to 2011, this study aims to trace back and date the origin of the cultivated genetic group suspected to have been founded by this major resistance breakdown event (Xhaard et al., 2011). Bayesian clustering analysis of microsatellite genotypes reveals that the cultivated group indeed dates back to the earliest virulent 7 samples in 1994. The pre-existing genetic group (gathering avirulent 7 individuals sampled in North-East France until 1997) vanishes with the emergence of the cultivated-virulent 7 group. Further analyses of the temporal samples shows evidence of a bottleneck in the earliest sample of 1994, but a quick recovery to genetic equilibrium and an increase in genetic diversity as the cultivated-virulent 7 group expanded. Our analysis highlights the importance of temporal sampling to unravel population turnover and temporal genetic characteristics of rapidly evolving organisms such as plant pathogens.

INTRODUCTION

How fast pathogen populations evolve is a central question in evolutionary biology (Williams, 2010; Galvani, 2003; McDonald and Linde, 2002). It determines the outcome of co-evolution with the host (Barret et al., 2009), the speed of evolutionary arm race (Tellier and Brown, 2011) and local adaptation (Gandon et al., 2008).

Plant domestication and agricultural conditions can imply drastic changes for the population biology of pathogens (Stukenbrock and McDonald, 2008) as shown for *Phytophthora infestans* (responsible for the potato late blight Goodwin et al., 1994; Montarry et al., 2010), *Magnaporthe oryzae* (rice blast; Couch et al., 2005; Saleh et al., 2012), and *Puccinia striiformis* (yellow wheat rust; Hovmoller and Justesen, 2007; Bahri et al., 2009; Ali et al., 2014). The high host densities and genetic uniformity of host populations resulting from cultivation practices create quasi-uniform environments that maintain pathogen population at large sizes and are conducive for disease development (Zhu et al., 2000). Its effects can be a speciation of pathogens driven by host specialization, as shown for *Rhynchosporium commune* and *Magnaporthe oryzae* (Zaffarano et al., 2008; Couch et al., 2005). Pathogens adapted to cultivated plants frequently exhibit strong differentiation and it often begins by a strong bottleneck characterized by a low level of genetic diversity compared to pathogens associated with wild hosts (Goodwin et al., 1994; Guérin et al., 2007; Stukenbrock and McDonald, 2008).

Plant resistance to fungal pathogens often relies on gene-for-gene interaction (Flor, 1971). In this model, the product of an avirulence (avr) gene in the pathogen is recognized by plants with a matching resistance (R) gene. Once the pathogen is recognized, a defense response is triggered (Jones and Dangl, 2006; de Wit, 2002). Conversely, pathogenicity factors, also known as effectors, contribute to the success of pathogen infection. For many cultivated species, resistances based on R genes have been favored in selection by plant breeders because they provide a complete control of the disease. However, despite its great potential efficiency, a gene-for-gene resistance may be broken down, as a single mutation in the pathogen can be sufficient to break the functionality of the avirulence gene, thereby restoring virulence (Schulze-Lefert et al., 1997). In the case of resistance breakdown, a virulence cost is often observed, characterized by a fitness decline associated with the emergence of a new virulence (Bahri et al., 2009).Escape from Host immunity by pathogens is frequently mediated by deletions or mutations in effector genes, which often show elevated levels of non-synonymous polymorphism as a result of their antagonistic co-evolution with the host (Stukenbrock and McDonald, 2008). In the case of a recent resistance breakdown, a loss of heterozygosity is observed (Gladieux et al., 2008; Burdon and Thrall, 2008), as a result of founder effect because of the initially limited number of individuals able to infect the hosts bearing the overcome resistance (Guérin et al., 2007).

Worldwide, *Melampsora* spp. (Basidiomycota, Pucciniales) are the most damaging pathogens of poplars *(*Steenackers et al., 1996*)*, and *Melampsora larici-populina* is the major problem in European poplar plantations (Frey et al., 2005). Poplars are particularly susceptible to *M. larici-populina* mostly because of their intensive monoclonal cultivation over several decades (Gérard et al., 2006). At the time of writing, eight qualitative resistances (R1 to R8) have been deployed in European plantations, and each has in turn been overcome by *M. larici-populin*a (Pinon and Frey, 2005). The most damaging resistance breakdown occurred when the widely planted resistance R7 was overcome and led to the invasion of France by virulent 7 *M. larici-populina* individuals (Pinon and Frey, 2005; Xhaard et al., 2011). Several poplar cultivars (or clones) bear the resistance R7, among which Beaupré was the most popular, accounting for up to 80% of the stems planted in northern France between 1980 and 1994 (Pinon and Frey, 1997). The first report of resistance R7 breakdown occurred in 1994 in Belgium and northern France (Pinon and Frey, 1997). In less than five years, virulent 7 individuals spread all over Western Europe and caused very destructive epidemics on all cultivated poplar stands (carrying or not resistance R7). It is noteworthy that unlike annual crop systems, such as wheat for which resistance phenotypes can be changed each year, poplar cultivation has a rotation of 15 to 25 years. Even if plantation of Beaupré rapidly collapsed after resistance R7 breakdown and the resulting massive epidemics, this cultivar still predominated poplar plantations of northern France for decades, thus exerting a sustained and strong selection pressure on *M. larici-populina* populations.

A previous population genetic study demonstrated that poplar rust populations have been greatly impacted by poplar cultivation (Xhaard et al., 2011). Current population structure of *M. larici-populina* in France is explained by only two major genetic groups. The most abundant gathers nearly all virulent 7 individuals and displays the hallmarks of a history of strong selection. It was thus named the cultivated group. This group predominates in northern France even on susceptible hosts (not carrying resistance R7). The second group is closer to genetic equilibrium and predominates only in southern France, especially in some refuges (like upstream the Durance valley in the Alps; Xhaard et al., 2012). Even if the Xhaard et al., (2011) study provides strong evidence that the resistance R7 breakdown was at the origin of the cultivated genetic group, it was not possible to date the foundation of this group, and therefore to provide a formal demonstration that it emerged at the same time than resistance R7 breakdown .

Using microsatellite genotyping of nearly 600 isolates from a historical collection (1992-2011) of *M. larici-populina* samples, this study aims to trace back the origin of the cultivated group, suspected to have been founded by the major resistance R7 breakdown event. Bayesian clustering and classical population analyses allowed to identify and to describe three genetic groups in our historical sampling. Two genetic clusters corresponded to the wild and cultivated groups already

described in Xhaard et al., (2011). Interestingly the cultivated group dated back to 1994, but we found no trace of it before the resistance R7 breakdown. Earlier samples consisted in a third genetic group which has disappeared from our samples since 1997 and was replaced by virulent 7 individuals. The historical collection also allowed to document a very quick recovery to genetic equilibrium of the cultivated group as well as an increase in genetic diversity following initial population bottleneck. Overall, this study highlights the strength of temporal population sampling for assessing the demographic history of parasite species characterized by rapid population turnover.

MATERIALS AND METHODS

Sampling strategy.

Sampling relies on a comprehensive collection of poplar rust isolates sampled since 1992 (Frey et al., unpublished). This collection was initially built with the aim of representing the diversity of rust populations, hence maximizing sampling locations and poplar cultivars of origin with the cost of reduced population sampling sizes. It was nonetheless possible to define, based on this collection, a coherent sampling scheme focusing on some locations regularly sampled through time (in particular in an experimental site at INRA Nancy). We enriched this sampling scheme with other sampled populations plus isolated individuals sampled at various locations, in order to provide a broad picture as we could of the temporal evolution in rust population genetic structure before and after the resistance 7 breakdown event. The sampling design encompasses 23 temporal samples (treated as populations) with each more than six individuals sampled at the same site and year, with a mean sample size of 14 individuals (Table 1). Those samples were mostly collected on susceptible poplar cultivars to reflect the composition of the population without selective filter at a given site and year as in Xhaard et al., (2011). Some individuals collected on R7 poplar cultivars just after the discovery of the resistance breakdown were added to examine further the genotypes of the very first virulent 7 individuals (Table 1). As a reference we added the genotypes of some individuals collected in 2009 and previously analyzed in Xhaard et al., (2011): the individuals sampled in Nancy, assigned to the cultivated group; and those from Prelles, located in the Alps (the South-East of France), in a valley free of virulent 7 individuals (Xhaard et al., 2012). Two additional samplings at Prelles at different years (2008 and 2011) were added to examine the temporal changes in genotype frequency at this location.

Individual isolation

Individuals were isolated from rust-infected poplar leaves collected in the field on various poplar cultivars bearing or not the resistance 7 (Table 1). For latter samples, one single sporulating lesion

of *M. larici-populina* (uredinia) per leaf was selected and grown on fresh leaf discs of *Populus* ×*euramericana* cv. Robusta as described by Barrès et al., (2008). The Robusta cultivar is susceptible to all rust isolates tested so far. Monouredinial isolation ensures that a single genotype per sampled leaf is isolated and multiplied. For older samplings, however, this protocol has not been strictly observed. A first batch of microsatellite genotyping was thus performed to check genotype purity of older isolates. In case of ambiguous genotypic profile, isolates were purified by picking a single spore from a sporulating lesion to start a new multiplication cycle on Robusta. After multiplication, leaf discs of Robusta were frozen and stored at -20°C until DNA extraction.

Virulence phenotype assessment

Most *M. larici-populina* isolates (Table 1) were inoculated on two poplar cultivars, one susceptible (*P.* ×*euramericana* cv. Robusta) and the second carrying the R7 rust resistance (*P.* ×*interamericana* cv. Beaupré), as described by Barrès et al., (2008). the phenotype of each individual was denoted Vir7 if both resistant and susceptible leaf discs were infected, and Avr7 if only the susceptible (control) leaf disc was infected. For each population, we calculated the proportion of Vir7 individuals.

Microsatellite genotyping

DNA was extracted using the BioSprint 96 DNA plant kit used in combination with the BioSprint 96 automated workstation (Qiagen®) following the BS96-DNA-plant protocol. Individuals were genotyped with a set of 21 microsatellite markers: MLP12 (Barrès et al., 2008), MLP49, MLP50, MLP54, MLP55, MLP56, MLP57, MLP58, MLP66, MLP68, MLP71, MLP77, MLP82, MLP83,MLP87, MLP91, MLP93, MLP94, MLP95, MLP96, MLP97 (Xhaard et al., 2009). Microsatellite markers were amplified by multiplex PCR, using the Multiplex PCR Kit (Qiagen®) as detailed in Xhaard et al., (2011). Three multiplex PCR were run, comprising 8, 5 and 8 loci, respectively (Xhaard et al., 2011; Table 2). PCR products from the three multiplex reactions (3 µL PCR1, 4 µL PCR2 and 5 µL PCR3) were pooled and loaded on an ABI 3730 Genetic Analyzer (Applied Biosystems). Genotyping was performed by the Genoscreen company. Fragments were sized with a LIZ-1200 size standard, and alleles were scored using GENEMAPPER 4.0 (Applied Biosystems). Individuals for which more than six loci failed to amplify were removed for further analyses.

Test for loci under selection.

We tested for signs of positive selection at each locus, using the F_{ST}-outlier approach (Lewontin and Krakauer, 1973) implemented in FDIST2 (Beaumont and Nichols, 1996). This method uses the

	Sampling on sensitive poplars				Sampling on R7 cultivars			
	Nancy	Orléans	NE France	Prelles	Grammont (Be)	Moy	NE France	Total
1992	17	6	1	0	0	0	0	24
1993	26	0	0	0	0	0	0	26
1994	12	0	0	0	8	0	14	34
1995	10	0	0	0	0	0	15	25
1997	13	0	0	0	0	0	2	15
1998	10	0	0	0	0	12	0	22
1999	8	0	0	0	0	0	0	8
2001	11	0	0	0	0	0	2	13
2002	3	0	0	0	0	0	0	3
2003	7	0	0	0	0	0	0	7
2004	6	11	0	0	0	0	0	17
2006	6	0	0	0	0	0	0	6
2007	14	0	0	0	0	0	0	14
2008	27	0	0	137	0	0	0	164
2009	18a	0	0	133a	0	0	0	151
2011	30	11	0	24	0	0	0	65
Total	218	28	1	294	8	12	33	594

Table 1: Summary of *M. larici-populina* isolates sampling. NE=North-East; Be=Belgium; a=including reference isolates from Xhaard et al., (2011).

expected distribution of F_{ST} vs. H_E to identify outlier loci that have excessively high (positive selection) or low (balancing selection) F_{ST} value compared with neutral expectations. We ran 100 000 simulations to compute the 95% and 99% confidence intervals of the neutral value distribution of F_{ST} values under the stepwise mutation model (SMM). Simulations were run on the 23 genetically homogeneous populations or at the genetic group level.

Population structure

We first investigated the population structure of the temporal samples of *M. larici-populina* using the Bayesian clustering method implemented in TESS (Chen et al., 2007; Durand et al., 2009). We used the model with admixture, which estimates putative multiple ancestry of individuals. The membership coefficient q to a given cluster is defined as the fraction of the genome that originates from this ancestral population. We used the conditional auto-regressive (CAR) Gaussian model of admixture with 100 000 MCMC iterations after the burn-in of 20 000 iterations. We performed five independent runs for each value of the maximal number of ancestral populations (K_{max}) ranging from 2 to 9. In order to choose the value of K_{max} that best suits the genetic data, TESS computes the Deviance Information Criterion (DIC) for each run (a procedure similar to the ΔK criterion traditionally used for structure [Evanno et al., 2005]). The DIC is an index of the model deviance penalized by its complexity (Spiegelhalter et al., 2002). The lower the DIC value, the higher the confident in the model. Moreover, the strength of the TESS method is that the number of clusters (K_{max}) assumed in a simulation run is the upper (not mandatory) number of ancestral populations. The hierarchical mixture algorithm implemented in TESS does not force individuals' genomes to be split in K_{max} populations, so that void clusters can appear if the population structure is already well explained by fewer clusters. In addition to the analysis of the decrease of DIC values, we also examined how the partition of individuals' genomes evolves when letting the number of clusters vary in order to select the most likely number of ancestral populations. Then, we fixed a threshold at q>0.8 to assign individuals to a given genetic group. This allowed to check the genetic homogeneity of the 23 pre-defined populations.

Population genetic analyses.

We computed summary statistics of genetic variability within groups identified using Bayesian clustering methods and within each genetically homogeneous population defined above. Expected (H_E) and observed (H_O) heterozygosities were calculated with GENETIX (4.05) (Belkhir et al., 1996, 2004). We reported the unbiased estimate of H_E calculated following Nei (1978). We calculated allelic richness (A_r) using FSTAT version 2.9.3 (Goudet, 1995) that implements a rarefaction procedure to account for differences in sample size. We reported these two estimates of

gene diversity as they can vary differently depending on the demographic regime (Cornuet and Luikart, 1996). Partition of genetic variability at the individual (F_{IS}) and population (F_{ST}) levels were estimated according to Weir and Cockerham (1984) using FSTAT. Linkage disequilibrium was estimated through the multilocus index of association corrected for the number of loci (rbarD) calculated using MULTILOCUS (Agapow and Burt, 2001). Significance of the departure from random association of alleles across loci was assessed by bootstrapping alleles among individuals 10 000 times.

Tests for demographic equilibrium.

Within population tests for mutation-drift equilibrium were performed with Bottleneck (Piry et al., 1999). This method relies on the fact that populations that have recently experienced a reduction in population size should exhibit larger values of gene diversity (i.e. heterozygosity level) than expected from the number of alleles at mutation–drift equilibrium (Cornuet and Luikart, 1996). Tests were performed assuming the two-phase mutational model (TPM), that better suits microsatellite markers. We allowed for 30% multistep changes in the TPM (default proportion). Analyses were performed with 2000 iterations of the coalescent process. Bottleneck probability was assessed using one-tailed Wilcoxon signed rank tests for heterozygosity excess (the most powerful test considering our limited number of markers [Luikart and Cornuet, 1999]).

Analysis of isolation by time

We tested the temporal evolution of allelic frequencies from Mantel tests between matrices of pairwise genetic distances between populations and of temporal distance, calculated as the number of years that separate the samples. We used the 'isolde' algorithm implemented in GENEPOP (Rousset, 2008), with $F_{ST}/(1-F_{ST})$ as estimate of genetic distances. Significance of regression slope was evaluated through 10 000 permutations.

RESULTS

Population structure analysis.

Bayesian clustering analysis of the historical collection of *M. larici-populina* samples indicated that our genetic data was best explained assuming 3 genetic groups. Deviance Information Criterion (DIC) showed a marked decrease between *K*max = 2 and 3, and then only a slight but regular decrease when increasing further the putative number of clusters (*K*max). This indicates that model prediction ability is best at *K*max = 3. Moreover, at *K*max = 4 and above, no individuals were assigned to the additional clusters. Assuming two clusters only, simulation runs gave two solutions. In most cases (4/5), TESS split the wild and cultivated groups as in Xhaard et al., (2011) (Fig. 1,

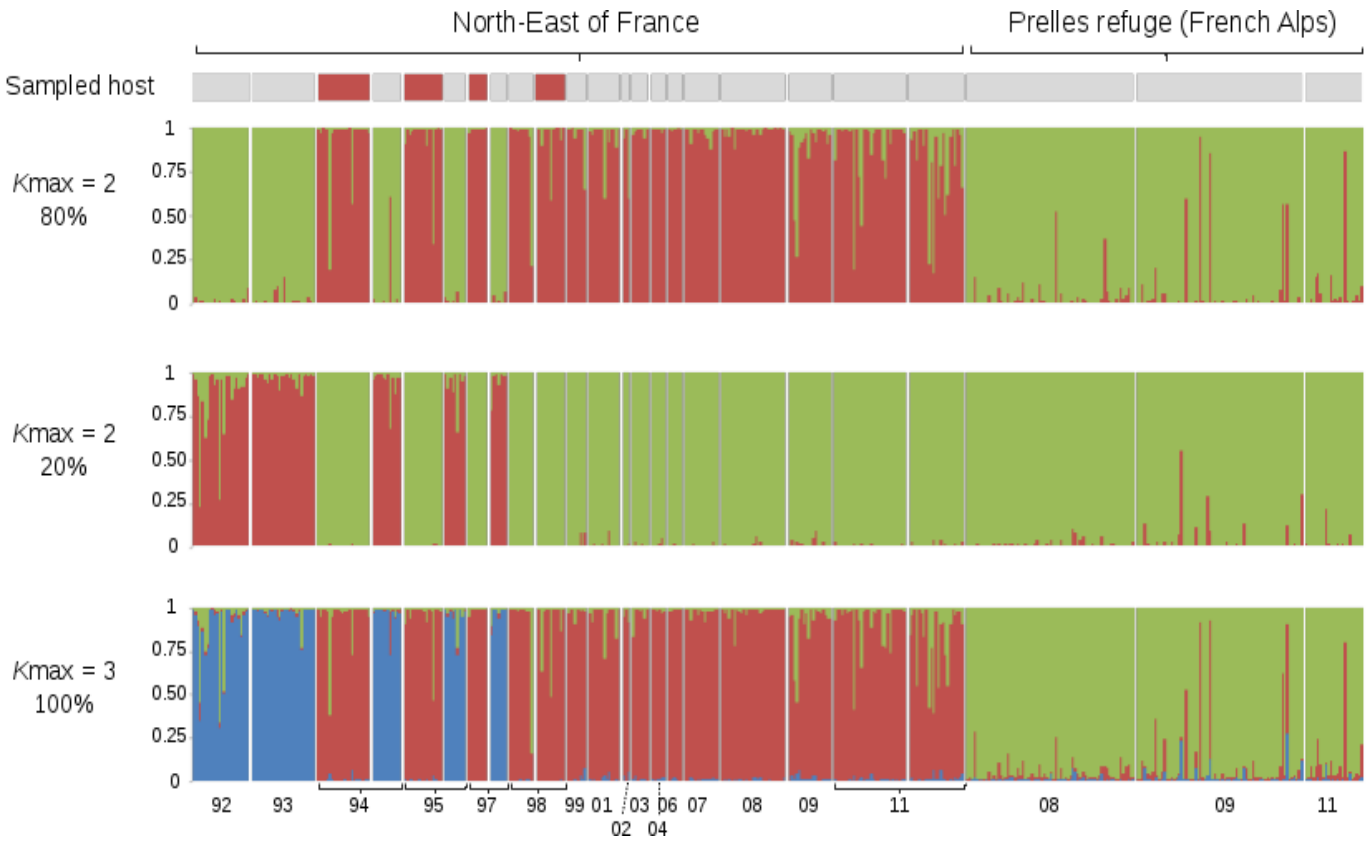


Figure 1: Barplots of membership probabilities for individual samples of the historical collection. We displayed the TESS results for Kmax = 2 (two solutions) and 3. Each vertical bar represents an individual whose genome is partitionned into up to 2 or 3 colored segments. Segment length is proportionnal to membership probability to each ancestral population. The type of host sampled (susceptible in grey, bearing resistance 7 in red) and the region of origin are reported above the barplots. Year of sampling (the last two digits) is indicated underneath.

*K*max = 2, first barplot). Interestingly, in those cases, early samples on resistance 7 poplar cultivars (since 1994) clustered with the cultivated group, but not the former and contemporaneous samples on susceptible poplars. In the alternative solution (Fig. 1, *K*max = 2, second barplot), the wild and cultivated groups clustered together, but the earlier samples on susceptible hosts formed a distinct cluster. At *K*max = 3 we found a unique solution that reconciled the two solutions of *K*max = 2 (Fig. 1, *K*max = 3). First, individuals sampled in the South of France at Prelles formed a first cluster, which corresponds to the wild group in Xhaard et al., (2011). Second, individuals sampled on resistance 7 cultivars from 1994 to 1998 grouped with individuals sampled since 1998 in the North-East of France. This cluster corresponds to the cultivated group described in Xhaard et al., (2011) and mostly consists in virulent 7 individuals (82% of the tested individuals bear the virulence 7, data not shown). Third, the very first samples of the historical collection (collected from 1992 to 1997 in the northern and eastern parts of France before the invasion of virulent 7 individuals) formed a third genetic group, that was not sampled, hence not studied, in Xhaard et al., (2011). No individual assigned to this group bears the virulence 7 (data not shown).

To distinguish between the two genetic groups of individuals sampled in the North-East of France in poplar cultivated areas, we decided to rename the first group (described in Xhaard et al., [2011)) the cultivated virulent group (CV) and the second group (newly discovered) the cultivated avirulent group (CA).

Focusing on the site of Nancy, we clearly observed a replacement of populations between 1997 and 1998, where CA individuals disappeared as the site became invaded by CV individuals. We observed asymmetric gene flow between groups. Only one CA individual sampled in 1994 showed some signs of hybridization with a CV individual. Conversely, no CV individual displayed trace of hybridization from CA genetic background. We found more pieces of evidence for gene flow between wild and cultivated groups (some individuals from both CA and CV had high membership probability to the wild group and some wild individuals had mixed ancestry to the three groups). Note however that this result has to be interpreted with caution and can simply result from erroneous assignments. Applying a threshold of $q>0.8$ on membership coefficients to assign individuals to the genetic groups led to the exclusion of only 21 individuals. More than 96% of the individuals were thus confidently assigned a genetic group. In the following, we focus on the 23 populations defined in Table 1 (more than six individuals sampled a given year, in a given region, sampled on a resistant 7 poplar tree). For the sake of genetic homogeneity, only the individuals confidently assigned to the dominant genetic group were retained for each population.

No sign of selection in our microsatellite marker set.

As the inclusion of selected markers can distort inference of demographic processes, we tested for

signs of positive selection at each locus, using the F_{ST}-outlier approach (Lewontin and Krakauer, 1973) implemented in FDIST2 (Beaumont and Nichols, 1996). Simulations were conducted considering either the three genetic groups or the 23 populations defined above. In both cases, all microsatellite loci (Fig. S3 and S4) fall within the 99% confidence intervals of neutral distribution of F_{ST} values under the stepwise mutation model (SMM). All loci can thus be considered as selectively neutral.

Genetic characteristics of the three groups

As expected, the wild group displayed the highest genetic diversity. Both allelic richness and heterozygosity levels were significantly higher than the cultivated groups (permutation test, $P < 0.001$). All wild populations were at genetic equilibrium (no deviation from Hardy-Weinberg proportions nor mutation drift equilibrium and lack of linkage disequilibrium; Table 2). There was no population differentiation (global $F_{ST} = 0.001$), although we observed a significant and positive global F_{IS} value.

Strikingly, the avirulent group showed the more pronounced deviations from genetic equilibrium (Table 2). Three (out of five) populations displayed positive and significant values of F_{IS} and rbarD (significant deviation from the Hardy-Weinberg equilibrium correlated with evidence for linkage disequilibrium). This was even more pronounced for the last population sampled from this group in 1997. Populations from this group also displayed the smallest values of genetic diversity (both *A*r and H_E values). Conversely, most populations of the virulent group were at genetic equilibrium. The notable exception is the very first sampling of this group in 1994. This population exhibited the highest value of F_{IS} and rbarD (significant deviation from the Hardy-Weinberg equilibrium and linkage disequilibrium). This population was also the only one to display a significant *P*-value for the Bottleneck test.

Temporal evolution of the cultivate virulent group

The cultivated virulent group presented the highest F_{ST} value (0.022) and the largest heterogeneity in various population genetic indices. We therefore subsequently tested for a temporal evolution of the genetic characteristics of this group. Indeed, we found a significant isolation by time (Fig. S5; $P < 0.05$). This change through time in allelic frequencies is accompanied by an increase in genetic diversity through years (Fig. 2). This was particularly obvious for *A*r, but the same trend can also be observed for H_E. Within a few years, the virulent group showed an increase of *A*r from 2.5 to 3 (respectively from 0.44 to 0.53 for H_E). Interestingly, this evolution in genetic diversity ranged from the genetic diversity level observed for the CA group (at the lower bound in earlier virulent samples) to the genetic diversity level found among populations of the wild group (upper bound

														Bottleneck			
														IAM		TPM	
		Year	N	Ar	He	Se(He)	Ho	Se(Ho)	Fis	CI (HW)	Fst	rbar D	CI (rbar D)	Pval	CI	Pval	CI
Avirulent group	92_Avr	1992	21	2,592	0,479	0,160	0,441	0,186	0,108	*		0,035	**	0,237		0,848	
	93_Avr	1993	26	2,422	0,436	0,185	0,413	0,212	0,074			0,013		0,156		0,785	
	94_Avr	1994	11	2,608	0,446	0,210	0,435	0,244	0,107	*		0,090	***	0,831		0,984	
	95_Avr	1995	9	2,455	0,455	0,185	0,372	0,229	0,080			-0,005		0,108		0,281	
	97_Avr	1997	7	2,557	0,438	0,196	0,443	0,246	0,137	**		0,101	**	0,507		0,912	
	All		**74**	**2,527**	**0,479**	**0,177**	**0,423**	**0,178**	**0,113**	***	**0,017**	**0,008**		**0,030**		**0,794**	
Virulent group	94_Vir	1994	17	2,560	0,489	0,206	0,449	0,194	0,249	***		0,067	***	0,002	*	0,027	
	95_Vir	1995	15	2,684	0,496	0,222	0,444	0,220	0,074			0,016		0,032		0,215	
	97_Vir	1997	8	2,585	0,450	0,222	0,446	0,258	0,066			-0,003		0,514		0,879	
	98A_Vir	1998	9	2,651	0,473	0,232	0,487	0,244	0,032			0,033		0,105		0,166	
	98B_Vir	1998	11	2,499	0,440	0,218	0,483	0,273	-0,044			0,006		0,226		0,671	
	99_Vir	1999	8	2,742	0,489	0,217	0,571	0,261	-0,103			0,012		0,057		0,435	
	01_Vir	2001	12	2,679	0,459	0,249	0,458	0,266	0,047			0,031		0,449		0,663	
	03_Vir	2003	7	2,774	0,476	0,232	0,457	0,225	0,119			0,008		0,194		0,493	
	04_Vir	2004	6	3,077	0,504	0,221	0,492	0,214	0,117			0,019		0,774		0,931	
	06_Vir	2006	6	2,492	0,438	0,242	0,484	0,329	-0,015			-0,015		0,045		0,162	
	07_Vir	2007	14	2,971	0,513	0,211	0,542	0,211	-0,019			0,002		0,160		0,774	
	08_Vir	2008	29	2,722	0,486	0,222	0,489	0,224	0,013			0,003		0,237		0,856	
	09_Vir	2009	16	2,836	0,515	0,176	0,470	0,185	0,122	*		-0,008		0,317		0,831	
	11A_Vir	2011	29	2,812	0,507	0,201	0,503	0,221	0,029			0,003		0,293		0,892	
	11B_Vir	2011	18	2,866	0,526	0,176	0,530	0,240	0,025			-0,007		0,160		0,620	
	All		**205**	**2,730**	**0,521**	**0,208**	**0,488**	**0,190**	**0,064**	***	**0,022**	**0,001**		**0,129**		**0,940**	
Wild group	08_W	2008	70	2,942	0,554	0,180	0,534	0,177	0,028			0,006		0,216		0,981	
	09_W	2009	63	2,960	0,543	0,186	0,536	0,193	0,022			-0,008		0,406		0,993	
	11_W	2011	23	3,037	0,557	0,175	0,540	0,178	0,054			-0,002		0,216		0,931	
	All		**156**	**2,980**	**0,556**	**0,179**	**0,535**	**0,173**	**0,030**	*	**0,001**	**0,000**		**0,247**		**0,994**	

Table 2: Genetic characteristics of the three gentic groups. SE : standard error ; HW : deviation for Hardy-Weinberg equilibirum ; CI : condidence interval with (*<0,01 ; **<0,001 ; ***<0,0001).

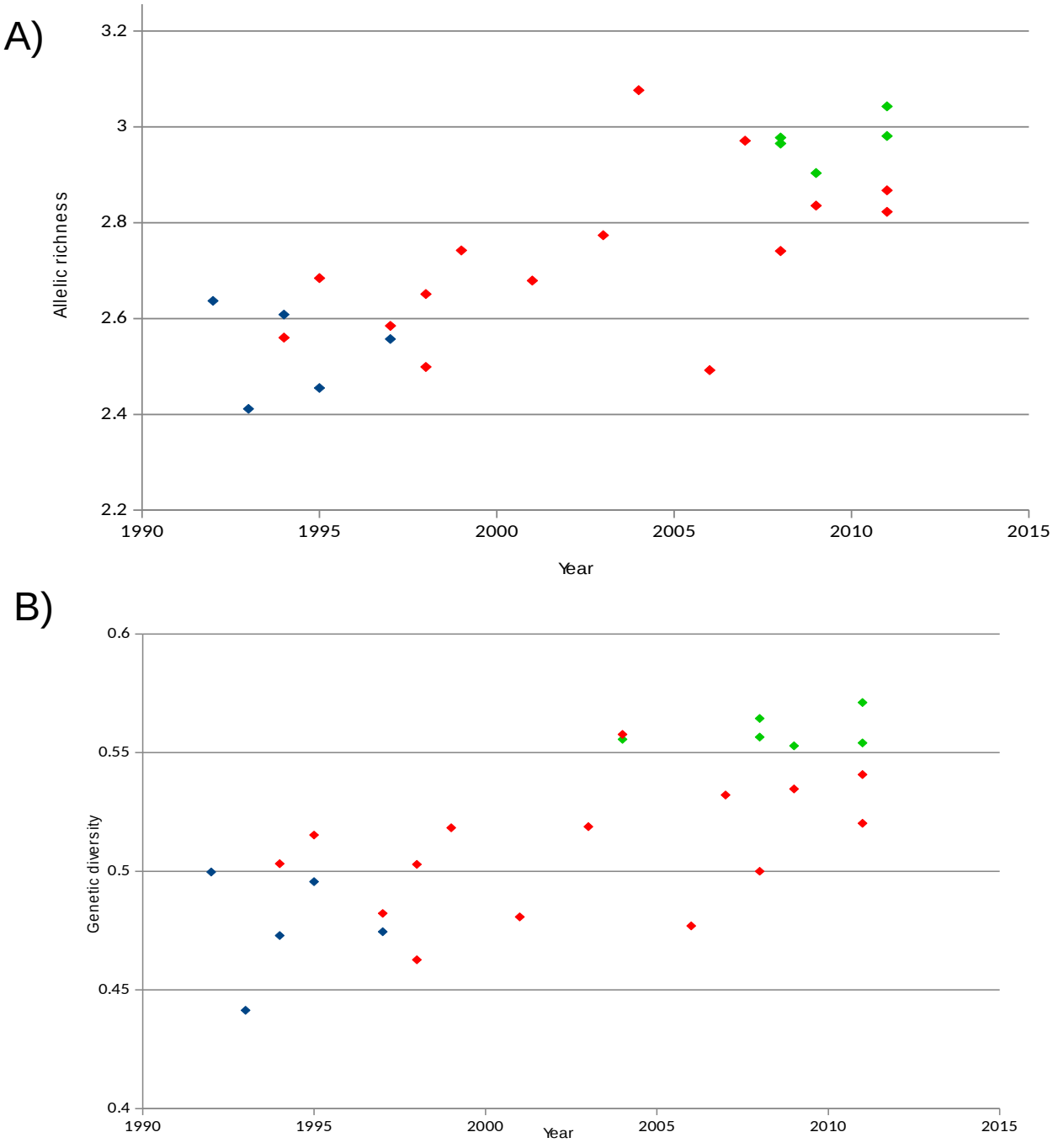


Figure 2: Evolution of allelic richness (A) and genetic diversity (B) since 1992 for the 23 populations. The three genetic groups described by TESS are represented in blue, red and green for the avirulent, virulent and wild group respectively.

reached since 2005).Moreover, the examination of this temporal pattern revealed that the increase in genetic diversity was only obvious after a time lag of four year (from 1994 to 1998 *A*r was stable and H_E displayed a slight decrease). Altogether, the variation of the population genetic indices provided several pieces of evidence that the virulent group experienced a bottleneck followed by strong demographic expansion.

Last we examined how the differentiation of the virulent group relatively to the wild and avirulent groups evolved over time. Overall, cultivated virulent populations were less and less differentiated from other populations through time (pairwise F_{ST} value against the avirulent group of 0.16 in 1994 and 0.11 in 2011, respectively 0.075 in 1994 and 0.045 in 2011 compared to the wild group; Fig 3). That could be the sign of ongoing gene flow between these populations during the demographic expansion following a founding effect. Interestingly, virulent populations are closer to the wild group but the decrease of F_{ST} is stronger against the cultivated avirulent group (slope of the regression of 7.17 and 2.55 for the avirulent and wild groups, respectively).

DISCUSSION

A historical collection of 594 isolates of *M. larici-populina* sampled in France and Belgium from 1992 to 2011 was used to study the effect of a major resistance breakdown, which was suspected to have shaped the present population genetic structure of this pathogenic fungus. The population genetic tools and the clustering analysis of genotyping data allowed us to identify and describe three genetic groups in our collection. These three groups are temporally and/or spatially separated. Contemporary samples revealed the same partition into two groups between northern and southern France, that was already highlighted in Xhaard et al., (2011): a dominance of the cultivated virulent 7 group in northern France, where Beaupré and other R7 cultivars were massively planted, and the presence of another genetic group in the south of France (in the Durance valley, in the French Alps). This second group overlaps with the native distribution area of *Populus nigra*, the wild host of *M. larici-populina* and was thus named the wild group (Xhaard et al., 2011). The Bayesian clustering analysis of the temporal sampling allowed us to consider two key elements for the understanding of the effect of the resistance R7 breakdown on the evolution of *M. larici-populina* population structure. First, we were able to link the date of resistance R7 breakdown with the date of emergence of the cultivated virulent 7 group, both observed in 1994. The very early samplings of virulent 7 individuals were grouped with contemporaneous samples in northern France (mostly virulent 7 phenotypes even if collected on susceptible hosts, data not shown). Second, Bayesian clustering analysis revealed that samples collected before resistance breakdown (and the detection of the first virulent 7 individuals) belong to a third genetic group, likely adapted to cultivated poplars not carrying the resistance R7.

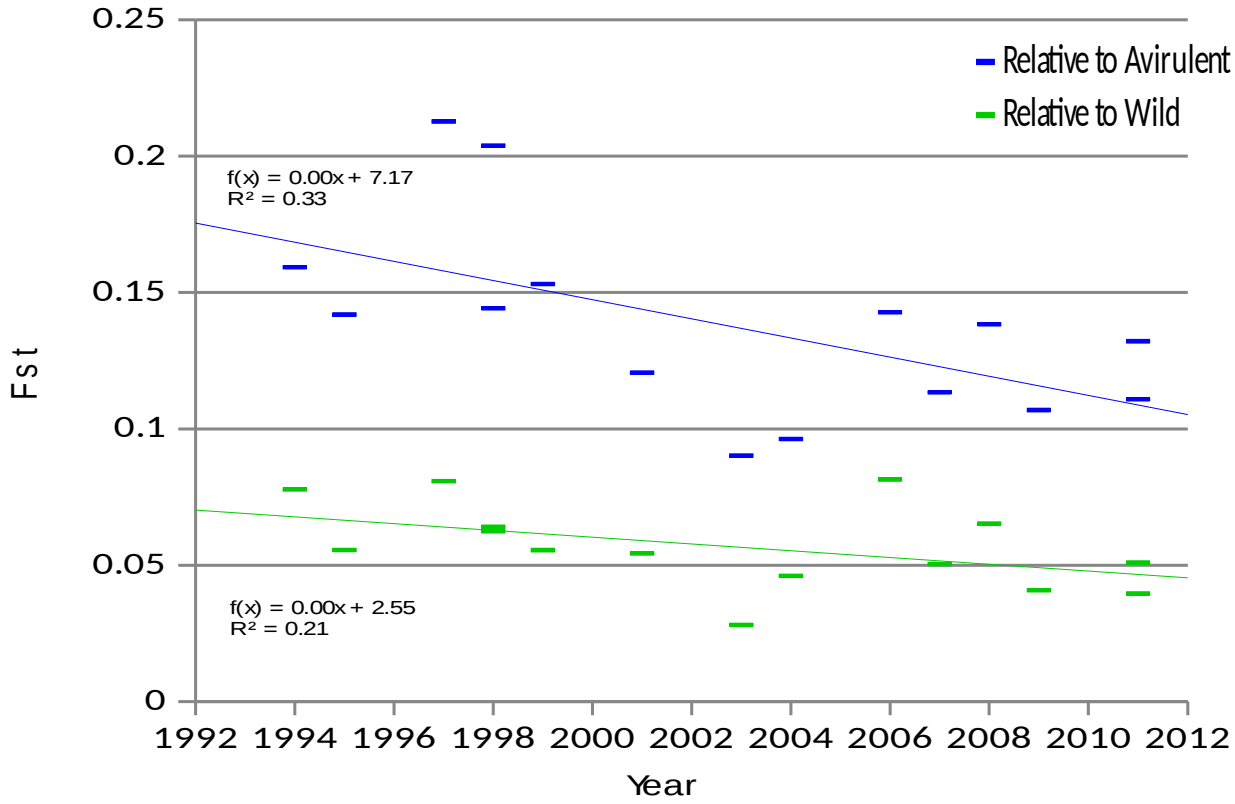


Figure 3: Temporal evolution between differentiation of the 14 Cultivated population against avirulent (blue) and aild (green) group. The F_{ST} values were computed for each Cultivated population against all Avirulent (blue) or Wild (green) sample. The two linear trendlines were added with their functions and R^2 in black box.

This genetic group vanished as cultivated virulent 7 genotypes invaded poplar stands. In less than five year we observed a complete replacement of *M. larici-populina* in northern France.

Isolates were genotyped with 21 microsatellite loci previously described (Xhaard et al., 2009). Microsatellites present several advantages compared to SNPs, including higher allelic richness, lower ascertainment bias, and amenability to large sample size at affordable cost (Schopen et al., 2008; Sun et al., 2009; Guichoux et al., 2011; Haasl and Payseur, 2011). Microsatellite loci could be subjected to selection and thus bias a population genetic study assuming neutrality (Haasl and Payseur, 2011). To avoid this pitfall, we check whether our marker set meets the neutral expectations by testing for signatures of positive selection on each locus, and no loci we found to be under selection, although three of them (Mlp54, 56 and 95) were previously described as under positive selection in the previous study of Xhaard et al., (2011). This can be explained by the sub-population structure that can strongly influence population genetic analysis (Karl et al., 2012) and mimic the effect of selection on these markers (Putman and Carbone, 2014). As the study of Xhaard et al., (2011) examined the genetic variability at many sites, the substructure (as found within the wild group) could have biased gene diversity estimates and generate false positives. Conversely, we performed here the selection test on very homogeneous genetic entities, which may explain this discrepancy and the lack of sign of positive selection with this data set.

Bayesian clustering analysis of the historical collection of *M. larici-populina* samples indicated that our genetic data were best explained assuming three genetic groups. Because this family of methods relies on MCMC algorithms, results from individual runs can vary and lead to different solutions (Gilbert et al., 2012). In order to take into account this variability we computed five runs for each *K*max value (ranging from *K*max=2 to *K*max=10). At *K*max=3 (best clustering solution), all runs converged to the same solutions with our three genetic groups. Conversely, assuming two clusters only (*K*max=2) leads to two solutions: either grouping CA and wild individuals together (major solution) or grouping CV and wild individuals. Inferring population structure when differentiation between populations is weak is a difficult analytic problem. For example the STRUCTURE models perform well at moderate levels of genetic differentiation ($0.02 < F_{ST} < 0.10$), but fails at lower values (Duchesne and Turgeon, 2012). Even if the F_{ST} value among populations was low (around 0.01 and 0.06 for W against CV in Xhaard et al., [2011]), the balanced sampling design and cluster sizes allowed a sound population structure to be obtained (more than 95% of sample confidentially assigned in one group with *K*max=3). We used the individual-based clustering method to directly infer migration by identifying individuals that belong to another, genetically distinct sub-population as implemented in TESS. The TESS individual-based clustering methods perform generally better than GENELAND, GENECLUST and STRUCTURE for inferring mixed ancestry (Chen et al., 2007). Interestingly, we found asymmetric gene flow between

groups with one CA individual sampled in 1994 showing some sign of hybridization with a CV individual but no CV individual displaying trace of CA genetic background. We found more pieces of evidence for gene flow between wild and cultivated groups. These results indicate potential strong gene flow between populations that could explain the low differentiation observed with F_{ST} (Waples and Gaggiotti, 2006).

The detailed examination of the variation of population genetic indices along the temporal samples of the cultivated virulent 7 group provided further insights on the demographic changes accompanying its emergence and spread. In particular we found that the earlier sample of 1994 displayed the hallmark of founder effect with a reduced genetic diversity (compared to later samples) and strong genetic disequilibria (significant deviations from Hardy-Weinberg proportions, expected heterozygosity and linkage across loci). Interestingly, the populations quickly returned to equilibrium with no significant deviation noticeable as early as 1995 (one year later). This fast disappearance of bottleneck signal can be an effect of the obligate annual sexual reproduction. The fact that *M. larici-populina* has to switch host to perform sexual reproduction increases the reshuffling of alleles among individuals (Barrès et al., 2008; Gilabert et al., 2009). Other biological system can retain the signatures of demographic event for longer periods, such as germ banking (seed or insect dormancy) which keeps signatures of ancient population decline for much longer time (Živković and Tellier, 2012). The obligate sexual reproduction (and host alternance) in *M. larici-populina* maximizes the effect of recombination and thus rapidly erases any trace of changes in population size. Another likely consequence of this genetic reshuffling is to uncouple most of the genome from the locus (or the loci) involved R7 breakdown.

The founder effect accompanying resistance breakdown should lead to a reduced genetic diversity compared to the population of origin (Guérin et al., 2007). Notably we found that initial CV samples displayed the same level of genetic diversity as the CA group that pre-existed at the same place. Based on the examination of virulence profiles, CV individuals likely originate from CA group as they share many virulence factors (except virulence 7), unlike wild individuals which display fewer virulence factors (data not shown). Furthermore, all CV individuals sampled in 1994 displayed the same combination of virulence factors and thus exhibit a much reduced diversity of virulence profile. It is thus puzzling not to observe the same pattern for neutral genetic diversity. This observation, together with the quick recovery to genetic equilibrium, points to the fact that resistance R7 might have been overcome several years before 1994.

A second originality in this temporal analysis came from the sustained increase in genetic diversity of the CV group through time. Interestingly the genetic diversity of this group reached the level observed for wild group. This variation went hand in hand with a decrease in differentiation level with wild group. This temporal evolution of genetic diversity could result from range

expansion of the cultivated virulent group, which increases the probability of gene mixing as CV individuals spread toward the south of France. We thus observed here a reverted profile of genetic diversity evolution compared to host tracking pathogens which display a reduction in diversity as it spread away from the center of domestication of the host plant (Gladieux et al., 2008; Saleh et al., 2012) as commonly found in phylogeographic studies of range expansion (Petit et al., 1997; Excoffier et al., 2009).

A perspective of this work would be to further assess the historical links between CA, CV and groups by testing alternative demographic scenarios using an ABC approach (Beaumont et al., 2002). Using this modeling framework, we would be able to test alternative hypotheses of the origin of the virulent 7 population (emergence from either wild or CA population, or another -not sampled- population) and to estimate key demographic parameters: effective population size, strength of the bottleneck caused by R7 breakdown, time of divergence of virulent 7 population, and rate of gene flow between populations. Such approach proved successful to unravel the demographic history of several cases fungal disease emergence worldwide (Barrès et al., 2012; Dutech et al., 2012; Gladieux et al., 2015).

CONCLUSION

While the rapid pace of pathogen evolution presents a major impediment to the development of sustainable crop protection strategies, it also provides interesting opportunities to document evolutionary processes at work. Nonetheless, the sampling design should cope with this accelerated evolution, as it challenges our view of studying population genetic structure. Typically, in our study, the CA group that predominated before the resistance R7 breakdown would have remained unnoticed without the temporal sampling (as in Xhaard et al., [2011]). Previous evolutionary studies highlighted the added value of temporal analysis of phenotypic variation to decipher host parasite co-evolution (Decaestecker et al., 2007) and the dynamics of adaptation (Blanquart and Gandon, 2013). Our population genetic analysis exemplifies further how temporal sampling could provide much more detailed insights on the demographic history of pathogen populations.

LITERATURE CITED

Agapow and Burt (2001) Indices of multilocus linkage disequilibrium. *Mol. Ecol. Notes* 1, pp101-102

Ali S, Shah SJ, Ibrahim M (2007) Assessment of wheat breeding lines for slow yellow rusting (*Puccinia striiformis West. tritici*). *Pak J Biol Sci*.10(19):3440-4

Bahri B, Leconte M, Ouffroukh A, De Vallavieille-Pope C, Enjalbert J (2009) Geographic limits of a clonal population of wheat yellow rust in the Mediterranean region. *Mol Ecol* 18: 4165–79

Barrès B, Halkett F, Dutech C, Andrieux A, Pinon J, Frey P (2008) Genetic structure of the poplar rust fungus *Melampsora larici-populina*: evidence for isolation by distance in Europe and recent founder effects overseas. *Infect Genet Evol*. 8(5):577-87

Barrès B, Carlier J, Seguin M, Fenouillet C, Cilas C, Ravigné V (2012) Understanding the recent colonization history of a plant pathogenic fungus using population genetic tools and Approximate Bayesian Computation. *Heredity.* 109(5):269-79

Barrett LG, Thrall PH, Dodds PN, van der Merwe M, Linde CC, Lawrence GJ, Burdon JJ (2009) Diversity and evolution of effector loci in natural populations of the plant pathogen *Melampsora lini*. *Mol Biol Evol.* 26(11):2499-513

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society B-Biological Sciences* 263: 1619–26

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*. 162(4):2025-35

Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (1996-2004) GENETIX 4.05, Logiciel Sous Windows TM Pour La Généique Des Populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5171, Université de Montpellier II, Montpellier (France)

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics.* 29: 1165–88

Blanquart F, Gandon S (2013) Time-shift experiments and patterns of adaptation across time and space. *Ecol Lett.* 16(1):31-8

Burdon JJ, Thrall PH (2008) Pathogen evolution across the agro-ecological interface: implications for disease management. *Evol. Appli.* 1: 57–65

Chen C, Durand E, Forbes F, François O (2007) Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Mol. Ecol. Notes*. 7:747–

756

Cornuet JM, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics.*144(4):2001-14

Couch BC, Fudal I, Lebrun M-H, Tharreau D, Valent B, Van Kym P, Nottéghem J-L and Kohn LM (2005) Origins of host-specific populations of the blast pathogen *Magnaporthe oryzae* in crop domestication with subsequent expansion of pandemic clones on rice and weeds of rice. *Genetics.* 170: 613–30

Decaestecker E, Gaba S, Raeymaekers JA, Stoks R, Van Kerckhoven L, Ebert D, De Meester L (2007) Host-parasite 'Red Queen' dynamics archived in pond sediment. *Nature.* 450(7171):870-3

De Mita S, Siol M (2012) EggLib: processing, analysis and simulation tools for population genetics and genomics. *BMC Genet.* 13:27

de Wit P.J (2002) Plant biology: on guard. *Nature* 416: 801–03

Duchesne P. Turgeon J (2012) FLOCK provides reliable solutions to the “number of populations” problem. *J. Hered.* 103:734–743

Durand E, Jay F, Gaggiotti OE, François O (2009) Spatial inference of admixture proportions and secondary contact zones. *Mol Biol Evol.* 26(9):1963-73

Dutech C, Barrès B, Bridier J, Robin C, Milgroom MG, Ravigné V (2012) The chestnut blight fungus world tour: successive introduction events from diverse origins in an invasive plant fungal pathogen. *Mol Ecol.* 21(16):3931-46

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol.* 14(8):2611-20

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*. 103(4):285-98

Flor H. H. (1971) Current status of the gene-for-gene concept. A*nnu. Rev. Phytopathol*. 9: 275–96

François S, Ancelet G. Guillot (2006) Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics* 174(2): 805-16

Galvani AP (2003) Immunity, antigenic heterogeneity, and aggregation of helminth parasites. *J Parasitol*. 89(2):232-41

Gandon S, Buckling A, Decaestecker E, Day T (2008) Host-parasite coevolution and patterns of adaptation across time and space. *J Evol Biol*. 21(6):1861-6.

Gérard P. R, Husson C, Pinon J, and Frey . (2006) Comparison of genetic and virulence diversity of *Melampsora larici-populina* populations on wild and cultivated poplar and influence of the alternate host. *Phytopathology* 96: 1027–36

Gilbert KJ, Andrew RL, Bock DG, Franklin MT, Kane NC, Moore J-S, et al. (2012) Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program STRUCTURE. *Mol. Ecol*. 21:4925–4930

Gladieux P, Zhang XG, Afoufa-Bastien D, Sanhueza RMV, Sbaghi M, Le Cam B (2008) On the origin and spread of the scab disease of apple: out of Central Asia. *PLoS ONE*. 3: e1455

Gladieux P, Feurtey A, Hood ME, Snirc A, Clavel J, Dutech C, Roy M, Giraud T (2015) The population biology of fungal invasions. *Mol Ecol.* 24(9):1969-86

Goodwin SB, Cohen BA, Fry WE (1994) Panglobal distribution of a single clonal lineage of the Irish potato famine fungus. *Proc Nat Acad Sci USA.* 91: 11591– 95

Goudet J (1995) FSTAT (Version 1.2): a computer program to calculate F-statistics. *Journal of Heredity.* 86(6): 485–86

Guérin F, Gladieux P, Le Cam B (2007) Origin and colonization history of newly virulent strains of the phytopathogenic fungus *Venturia inaequalis*. *Fungal Genetics and Biology*. 44: 284–292

Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger P, Lepais O, et al. (2011) Current trends in microsatellite genotyping. *Mol. Ecol. Resour*. 11:591–611.

Haasl RJ, Payseur BA (2011) Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. *Heredity.*;106:158–171

Haasl RJ,. Payseur BA (2013) Microsatellites as targets of natural selection. *Mol. Biol. Evol.* 30:285–298

Hovmøller MS, Justesen AF (2007) Rates of evolution of avirulence phenotypes and DNA markers in a northwest European population of *Puccinia striiformis f. sp.tritici. Mol Ecol* 16(21):4637-47

Huang J, Si W, Deng Q, Li P, Yang S (2014) Rapid evolution of avirulence genes in rice blast fungus Magnaporthe oryzae. *BMC Genet.* 15(45)

Jones J.D and Dangl J.L (2006) The plant immune system. *Nature.* 16: 323-29 doi:10.1038/nature05286

Karl SA, Toonen RJ, Grant WS. Bowen BW (2012) Common misconceptions in molecular ecology:

echoes of the modern synthesis. *Mol. Ecol*. 21:4171–4189

Leroy T, Lemaire C, Dunemann F, Le Cam B (2013) The genetic structure of a *Venturia inaequalis* population in a heterogeneous host population composed of different Malus species. *BMC Evol Biol*. 13(64)

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*. 74(1):175-95

Luikart G, Cornuet JM (1999) Estimating the effective number of breeders from heterozygote excess in progeny. *Genetics*.151(3):1211-6

McDonald BA, Linde C (2002) Pathogen population genetics, evolutionary potential, and durable resistance. *Annu Rev Phytopathol*. 40:349-79

Montarry J, Hamelin FM, Glais I, Corbi R, Andrivon D.(2010) Fitness costs associated with unnecessary virulence factors and life history traits: evolutionary insights from the potato late blight pathogen *Phytophthora infestans*. *BMC Evol Biol*

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*. 89(3):583-90

Petit RJ, Pineau E, Demesure B, Bacilieri R, Ducousso A, Kremer A (1997) Chloroplast DNA footprints of postglacial recolonization by oaks. *Proc Natl Acad Sci* 94(18):9996-10001

Pinon, J., and Frey, P (1997). Structure of *Melampsora larici-populina* populations on wild and cultivated poplar. Eur. J. Plant Pathol. *103*, 159–173

Pinon, J. and Frey, P (2005). Interactions between poplar clones and *Melampsora* populations and their implications for breeding for durable resistance. In: Rust Dis. Willow Poplar. Pei, M. H., McCracken, A. R., pp. 139–154

Piry S, Luikart G, Cornuet JM (1999) BOTTLENECK: a computer program for detecting recent reductions in the effective population size using allele frequency data. *Journal of Heredity.* 90: 502–03

Putman AI, Carbone I (2014) Challenges in analysis and interpretation of microsatellite data for population genetic studies. *Ecology and Evolution* 4(22):4399-4428

Rousset F (2008) GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Mol. Ecol. Res*. 8: 103–06

Saleh D, Xu P, Shen Y, Li C, Adreit H, Milazzo J, Ravigné V, Bazin E, Nottéghem JL, Fournier E, Tharreau D (2012) Sex at the origin: an Asian population of the rice blast fungus *Magnaporthe oryzae* reproduces sexually. *Mol Ecol.* 21(6):1330-44

Schulze-Lefert P, Peterhaensel C, Freialdenhoven A (1997). Mutation analysis for the dissection of resistance. In: Crute I.R, Holub E.B, Burdon J.J. (Eds.) The Gene-for-Gene Relationship in Plant–Parasite Interactions. *CAB International, New York* pp. 45–63

Schopen GCB, Bovenhuis H, Visker MHPW. van Arendonk JAM (2008) Comparison of information content for microsatellites and SNPs in poultry and cattle. *Anim. Genet.* 39:451–453

Spiegelhalter, S. D., Best, N. G., Carlin, B. P., and Linde, A. V. D (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 64(4): 583–639

Steenackers, J., Steenackers, M., Steenackers, V., and Stevens, M (1996) Poplar diseases, consequences on growth and wood quality. *Biomass Bioenerg*. 10: 267–74

Stukenbrock, E. H., and McDonald, B. A (2008). Population genetics of fungal and oomycete effectors involved in gene-for-gene interactions. *Mol. Plant Microbe Interact*. 22: 371–80

Stukenbrock EH, Bataillon T (2012) A population genomics perspective on the emergence and adaptation of new plant pathogens in agro-ecosystems. *PloS Pathog*. 8(9):e1002893

Sun JX, Mullikin JC, Patterson N. Reich DE (2009) Microsatellites are molecular clocks that support accurate inferences about history. *Mol. Biol. Evol.* 26:1017–1027

Tellier A, Brown JK (2011) The influence of perenniality and seed banks on polymorphism in plant-parasite interactions. *Am Nat*.174(6):769-79

Waples RS. Gaggiotti O (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Mol. Ecol.* 15:1419–1439

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38: 1358–70

Williams PD.(2010) Darwinian interventions: taming pathogens through evolutionary ecology. *Trends Parasitol.* 26(2):83-92

Xhaard C, Andrieux A, Halkett F, Frey P (2009) Characterization of 41 microsatellite loci developed from the genome sequence of the poplar rust fungus, *Melampsora larici-populina*. *Conservation Genetics Res.* 1: 21–25

Xhaard C, Fabre B, Andrieux A, Gladieux P, Barrès B, Frey P, Halkett F (2011) The genetic structure of the plant pathogenic fungus M*elampsora larici-populina* on its wild host is extensively impacted by host domestication. *Mol Ecol.* 20(13):2739-55

Xhaard C, Barrès B, Andrieux A, Bousset L, Halkett F, Frey P (2012) Disentangling the genetic origins of a plant pathogen during disease spread using an original molecular epidemiology approach. *Mol Ecol.* 21(10):2383-98

Zaffarano PL, McDonald BA, Linde CC (2008) Rapid speciation following recent host shifts in the plant pathogenic fungus *Rhyncosporium*. *Evolution* 62: 1418–36

Živković D, Tellier A (2012) Germ banks affect the inference of past demographic events. *Mol Ecol*. 21(22):5434-46.

Zhu Y, Chen H, Fan J, Wang Y, Li Y, Chen J, Fan J, Yang S, Hu L, Leung H, Mew TW, Teng PS, Wang Z, Mundt CC.(2009) Genetic diversity and disease control in rice. *Nature*. 406(6797):718-22

Supplemental material

Figure. S1: Decrease of the Deviance Information Criterion (DIC) with the assumed number of clusters (*K*max) given by the TESS analysis. The diamonds corresponds to the mean DIC value (over 5 runs) and the bars indicates standard deviation.

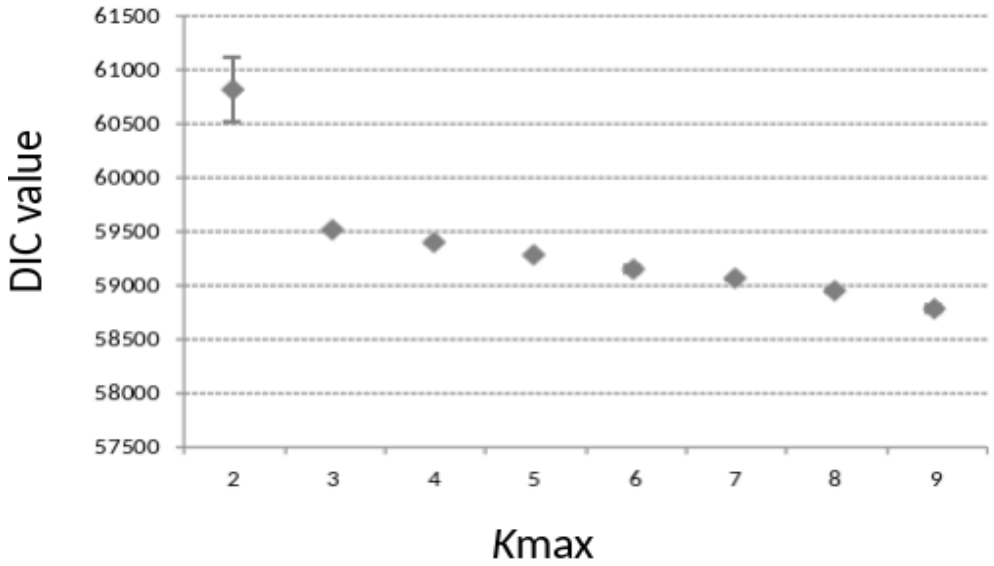


Figure S2: Evolution of the mean membership coefficient (mean *q*) to the cultivated avirulent group (CA) in Nancy from 1992 to 2011.

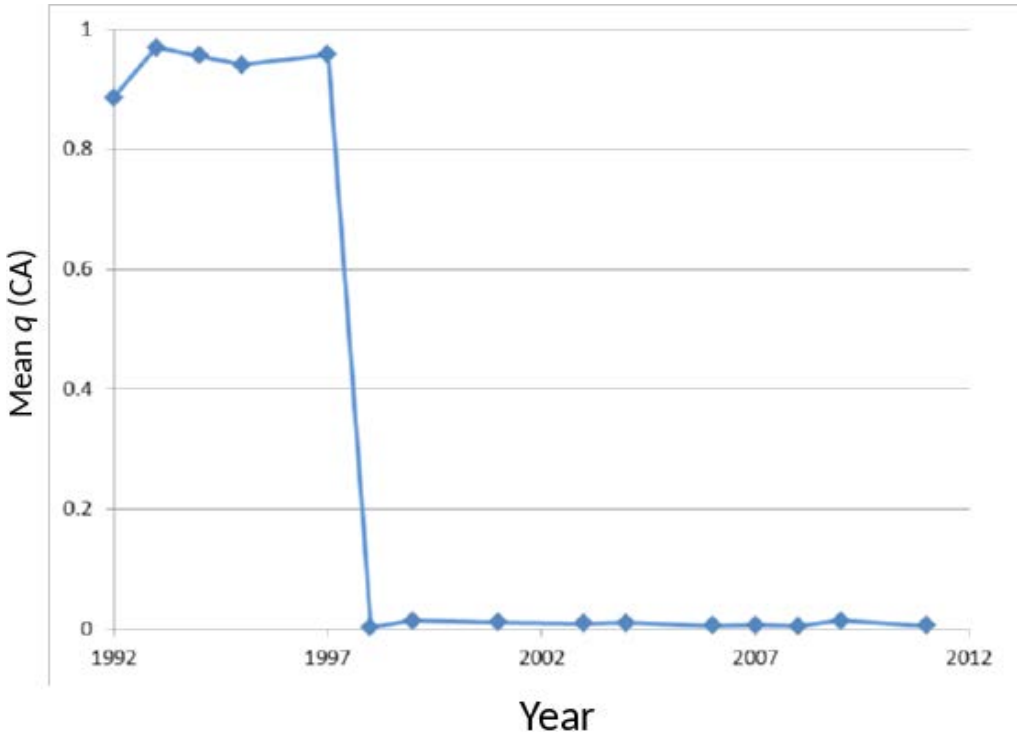


Figure S3: Test for selection for 21 microsatellitte loci with 23 populations. Observed F_{ST} values are plotted against heterozygosity levels for each locus. Solid and dashed lines denote median and 95% confidence interval, respectively, of the simulated distributions under the stepwise mutation model (SMM). Solid grey line indicates the 99% confidence interval. All loci were found to be consistent with neutral expectations (black diamonds).

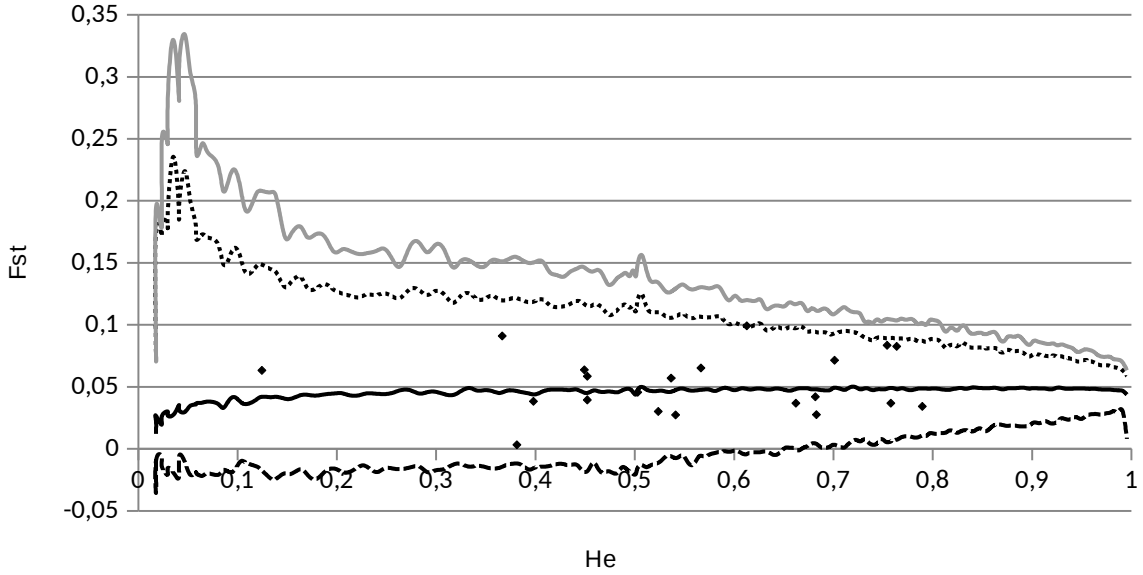


Figure S4: Test for selection for 21 microsatellitte loci with 3 populations. Observed F_{ST} values are plotted against heterozygosity levels for each locus. Solid and dashed lines denote median and 95% confidence interval, respectively, of the simulated distributions under the stepwise mutation model (SMM). Solid grey line indicates the 99% confidence interval. All loci were found to be consistent with neutral expectations (black diamonds).

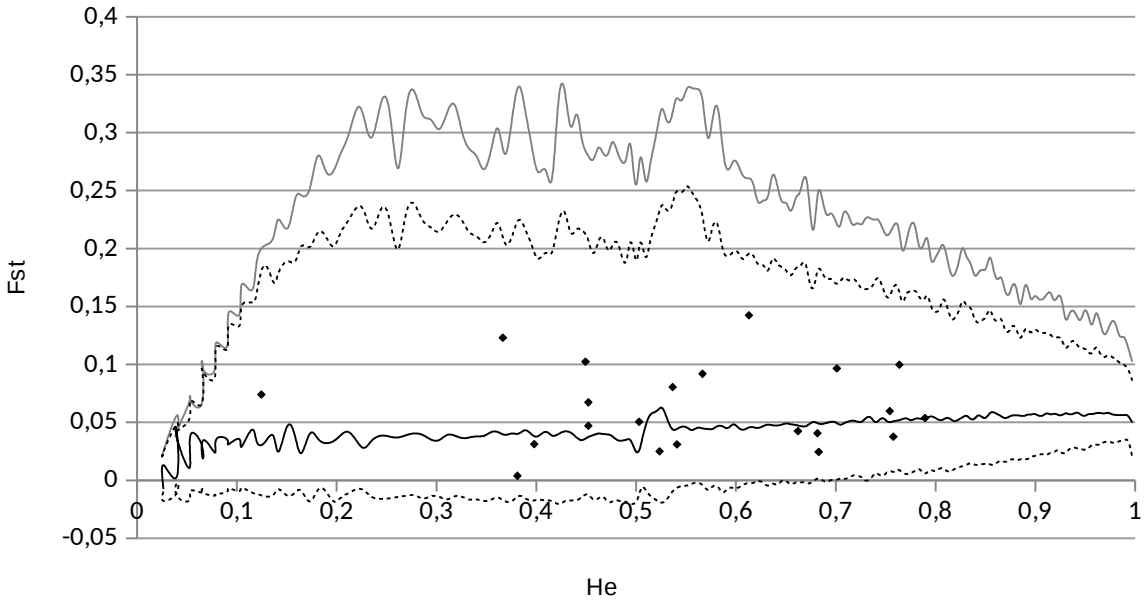
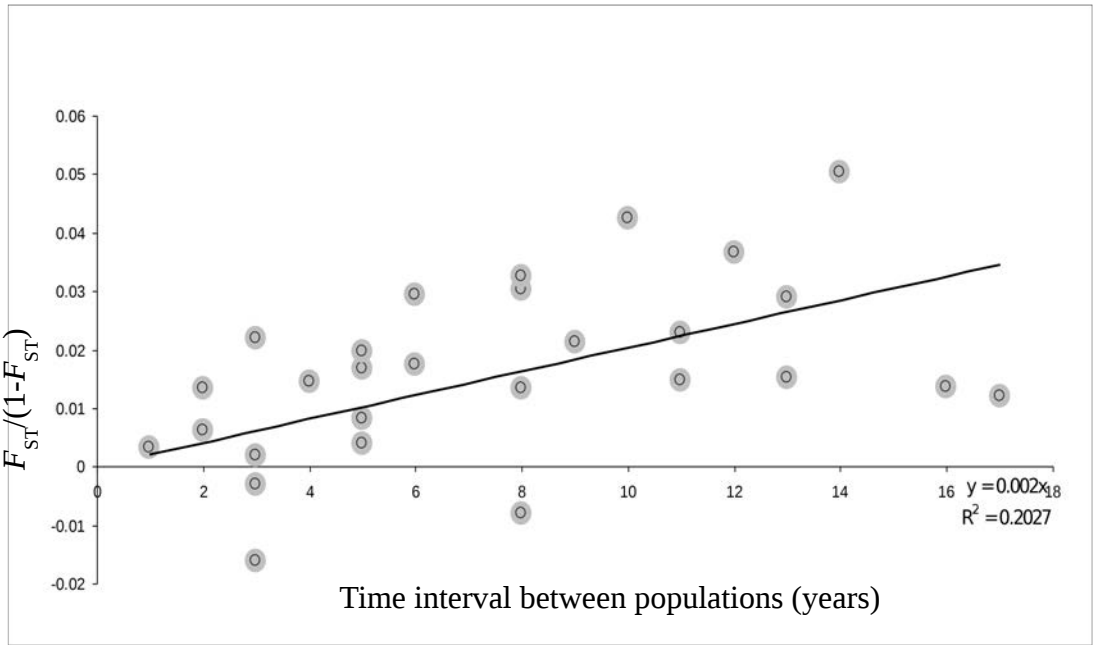


Figure S5: Isolation by time of the cultivated virulent 7 group since 1994 represented by linear regression ($P < 0.05$). The X axis represents the time interval between two populations and the Y axis a standardized measure of F_{ST} between pairs of populations.



Chapitre 3-bis

Vers l'obtention d'un scénario démographique

Chapitre 3-bis - Vers l'obtention d'un scénario démographique

1.Introduction

Le Chapitre 3 nous a permis de décrypter l'impact fort du contournement de la résistance 7 sur l'histoire démographique de *M. larici-populina* (Persoons et al., in prep). Trois groupes génétiques ont été décrits dont deux directement impactés par le contournement, les avirulents 7 semblent avoir disparu du territoire Français à partir de 1997, vraisemblablement remplacés par les virulents 7 apparus en France en 1994 (année pressentie du contournement).

Dans le but de comprendre les liens historiques entre les populations, une première étude d'inférence démographique basée sur la coalescence a été réalisée à l'aide du logiciel DIY ABC (Do It Yourself Approximate Bayesian Computations, version 0.7, (Cornuet et al., 2008). Par cette analyse nous souhaitions confirmer le goulot d'étranglement et l'expansion démographique des virulents 7. De plus, on s'est intéressé à l'origine de ce groupe génétique : Est-il apparu au sein des sauvages qui présentent une grande diversité génétique (Xhaard et al., 2011) ou au sein des avirulents 7 avec lesquels il a co-existé dans le Nord-Est de la France entre 1994 et 1997 (Persoons et al., in prep) ?

Cette étude préliminaire de coalescence est basée sur le génotype de 119 isolats du Chapitre 3 regroupés en 4 populations. Étant donné qu'une l'analyse démographique plus complète a été réalisée ensuite avec les données génomique (Chapitre 4), cet ajout au chapitre 3 ne présentera que les résultats principaux acquis à l'aide du logiciel DIY ABC sur ce sous-échantillon des données microsatellites.

2. Matériel et méthodes

Dans le but d'inférer les paramètres décrivant l'histoire de nos populations l'ABC a été utilisée à l'aide du programme DIY ABC (Do It Yourself Approximate Bayesian Computations, version 0.7, (Cornuet et al., 2008). Il s'agit une approche bayésienne dans laquelle sont estimées les probabilités relatives de différents modèles fixés *a priori* ainsi que les distributions des valeurs des paramètres. Pour se faire nous avons choisi 4 populations clefs à partir des résultats du Chapitre 3 (Tableau 1). Ensuite nous avons établis quatre scénarios démographiques (Figure 1) avec comme distribution *a priori* des paramètres :

Ayant très peu d'information, nous avons choisi une distribution uniforme bornée entre 100 et 10 000 pour les 3 tailles de populations

Année	Groupe génétique	Nombres d'isolats	Localisation (départements)
1992-1994	Avirulents 7	21	Amance (54)
1994-1995	Virulents 7	27	Nord-Est (02,45,55,59,62)
2009	Sauvage	40	Prelles (05)
2011	Virulents 7	31	Amance (54)

Tableau 1 : Échantillonnage utilisé dans l'étude. La première population des Virulents 7 regroupe des isolats collectés à plusieurs endroits (contingence due à la mise en collection à ces dates).

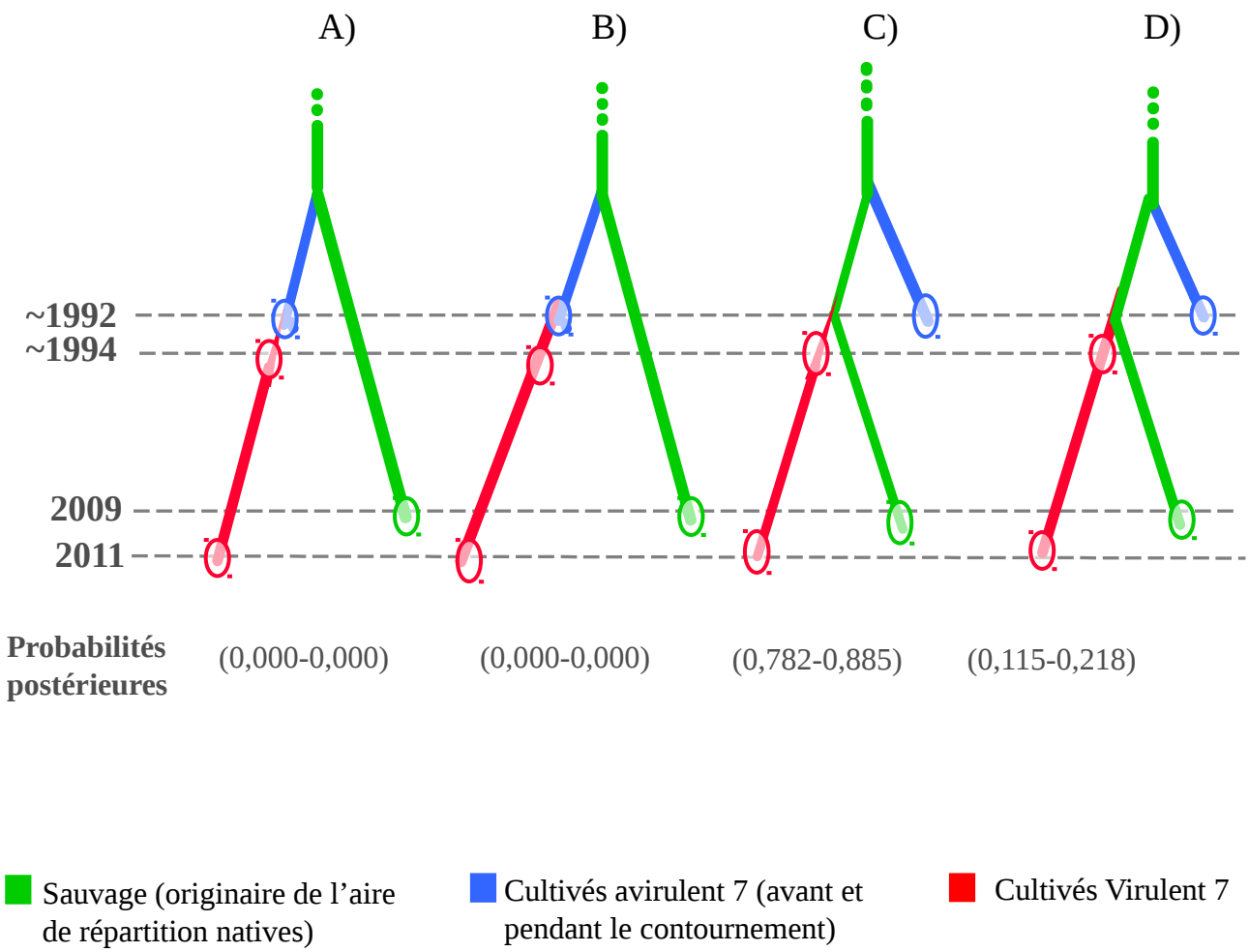


Figure 1 : Scénarios de coalescence pour la fondation du groupe cultivé virulent 7. Pour l'ensemble de ces scénarios, les pré-requis sont que le groupe sauvage est le groupe ancestral et qu'il n'y a pas de population « fantôme » qui aurait disparue sans être échantillonnée. Dans ce cadre, les scénarios A et C considèrent un bottleneck à la naissance du groupe cultivé virulent 7 (représenté par un pincement du trait). Les scénarios A et B prévoient que le groupe cultivé virulent 7 descend du groupe cultivé avirulent contrairement aux scenarios C et D qui le font descendre du groupe sauvage. Les probabilités postérieures sont calculées par le logiciel DIY ABC (Cornuet et al., 2010) après 1 000 000 de simulations.

- Le temps de fusion entre les populations sauvage et avirulente est également tiré dans un prior peu informatif (Uniforme entre 50 et 1000 générations, une génération sexuée équivalent à une année dans le cas de notre modèle biologique).
- Par contre nous avons resserré les distributions *a priori* pour le temps de fusion de la population virulente (Log Uniforme entre 18 et 49 générations, 18 correspondant à l'écart entre les deux dates d'échantillonnage du groupe virulent, soit une émergence en 1993) ; et pour les deux scénarios incluant un goulot d'étranglement, la taille de la population virulente lors de sa fondation (Log Uniforme entre 2 et 50 générations) et la durée du goulot d'étranglement dont on a vu en début de chapitre qu'elle était certainement très faible (Uniforme entre 1 et 10 générations).

Afin de se conformer au modèle mutationnel de type SMM implémenté dans DIY ABC, nous n'avons retenu que les marqueurs microsatellites présentant un motif de répétition standard (19 locus : Mlp_93, Mlp_91, Mlp_55, Mlp_83, Mlp_57, Mlp_82b, Mlp_66, Mlp_12, Mlp_58, Mlp_50, Mlp_56, Mlp_87, Mlp_49, Mlp_96, Mlp_94, Mlp_68b, Mlp_97, Mlp_71, Mlp_95).

Pour l'algorithme ABC nous avons retenu cinq statistiques descriptives : le nombre d'allèles et l'hétérozygotie pour les quatre populations échantillonnées et la valeur du F_{ST}, la variation des tailles d'allèle et la distance en allèles partagés pour les six paires de populations.

Les 4 scénarios ont ensuite été simulés avec 1 000 000 d'itérations par scénarios. Le meilleur modèle a été sélectionné par régression logistique, générant une probabilité postérieure permettant de connaitre parmi les scénarios simulés celui présentant le niveau d'adéquation le plus important avec nos données (Fagundes et al., 2007). Nous avons vérifié par ACP que les simulations produisaient un nuage de valeurs simulées assez dense autour de nos données observées. Les distributions *a posteriori* des paramètres du modèle sélectionné ont été générées en retenant les 5% des simulations donnant des valeurs de statistiques descriptives les plus proches de nos données.

2. Résultats et discussion

L'échantillonnage utilisé dans cette analyse est le génotype de 119 isolats répartis en 4 populations clefs. Suite aux simulations des quatre scénarios, le scénario qui présente les meilleures probabilités (Intervalle de confiance à 95% 0,782 à 0,885) est le scénario C qui tient compte d'un goulot d'étranglement et postule l'origine du groupe cultivé virulent 7 depuis le groupe sauvage (figure 1). Le deuxième scénario le plus probable est le scénario D (sans goulot d'étranglement). Une analyse plus avancée des résultats de simulation montre qu'il y a peu de confusion possible entre ces quatre types de scénario (plus de 90% des simulations générées sous le scénario C sont bien assignées comme tel). Les deux scénarios A et B (les virulents 7 dérivant des avirulents) présentent une

probabilité nulle. Le groupe des virulents 7 serait dans apparu dans un fond génétique sauvage et non des avirulent 7 comme leur proximité géographique, phénotypique et écologique (hôtes cultivés) aurait pu le laisser penser. Ainsi la question est de savoir pourquoi la mutation vers la virulence 7 est apparues chez les sauvages et non chez les cultivés avirulents. Comme ce groupe sauvage est plus diversifié génétiquement, il est possible que cette ou ces mutations soi(en)t apparue(s) dans ce groupe puis que par migration il se soit retrouvé en contact avec les Beauprés virulent 7, rendant les individus porteurs de cette virulence aptes à envahir ces peupleraies. Si tel est le cas l'identification des déterminants de cette virulence 7 devrait permettre d'en retrouver les traces dans ce groupe sauvage.

De nombreuses limites existent concernant cette analyse. Tout d'abord le logiciel ne prend pas en compte les flux de gène entre les populations ce qui constitue une limite d'autant plus forte de ce modèle, que l'analyse des données microsatellite montre de forts flux de gènes surtout entre groups sauvage et cultivés. Cette dernière analyse, fournissant même une estimation basse de ce flux de gènes puisque les populations virulente et sauvage ont été échantillonnées dans des régions très différents et pour la population sauvage dans une vallée refuge des alpes (en amont de la Durance Xhaard et al., 2012). Par ailleurs, le logiciel DIY ABC sort toujours un scénario favori même si aucun ne correspond aux données, ce qui peut conduire à des conclusions erronées. Pour toutes ces raisons, nous avons décidé de ne pas poursuivre plus avant de type d'analyse avec ce logiciel. Une approche d'inférence *ad hoc* sera menée sur les données de génomique des populations.

Chapitre 4

Conséquences démographiques et génomiques d'un évènement majeur de sélection chez l'agent de la rouille du peuplier *Melampsora larici-populina*

Chapitre 4 - Conséquences démographiques et génomiques d'un évènement majeur de sélection chez l'agent de la rouille du peuplier *Melampsora larici-populina*

1. Introduction

Dans les chapitres précédents nous avons décrit l'effet du contournement de la résistance 7 sur les populations de *M. larici-populina* (chapitre 3) et nous avons tenté d'obtenir un scénario démographique (chapitre 3-bis) avec des données de génotypage. L'objectif de ce chapitre est d'étudier l'évènement de sélection qui a accompagné le contournement de la résistance 7 par une étude de génomique des populations.

Pour cela 86 isolats de *M. larici-populina* répartis en quatre populations clefs issues d'échantillonnages ayant eu lieu de part et d'autre du contournement ont été sélectionnés parmi ceux du chapitre 3 et leur génome a été séquencé. Le polymorphisme a été étudié après alignement sur la nouvelle version du génome de référence, assemblé en 18 groupes de liaisons (pseudo-chromosomes) et plus d'un millions de SNP ont été mis à jour. Le déséquilibre de liaison a été mesuré sur l'ensemble des données et séparément par population nous permettant d'observer une décroissance du DL variant en intensité en fonction de la population. Nous avons mesuré le spectre de fréquence allélique global (SFS) et le spectre joint population contre population (JFS) ainsi que les valeurs de plusieurs statistiques mesurant l'écart du patron de diversité génétique par rapport aux attendus neutres.

Deux scénarios démographiques ont été conçus et utilisés dans le cadre de simulations de coalescence à l'aide d'EggLib (De Mita and Siol, 2012). Le choix du modèle le plus vraisemblable et l'estimation des paramètres ont été réalisés par ABC en utilisant comme statistiques résumantes les SFS et JFS, nous permettant d'obtenir un scénario démographique fiable avec les caractéristiques suivantes : l'ancêtre commun du groupe des virulents 7 et des avirulents 7 serait plus récent (~ 2100 ans) que celui des virulents 7 et des sauvages (~2300 ans). Le groupe des virulents 7 aurait subit des flux de gènes conséquents avec les avirulents 7 et les sauvages et un goulot d'étranglement de force modérée au moment du contournement de la résistance 7.

Des scans génomiques ont ensuite été réalisés en mesurant le D de Tajima et le F_{ST} contrastant les groupes virulent 7 et avirulent 7. Des intervalles de confiance pour ces indices ont été obtenus grâce au modèle démographique et une recherche des régions génomiques présentant un D de Tajima et un F_{ST} significativement en dehors de l'intervalle a été conduite. Cela a permis de

détecter 20 régions génomiques présentant un patron de sélection potentiellement lié au contournement de la résistance 7. Ces régions contiennent 14 gènes considérés dès lors comme des gènes candidats associés à la virulence 7.

Ce chapitre se présente sous la forme d’un article en préparation.

2. Article n°3 : Demographic and genomic consequences of a major event of adaptation in the pathogenic fungus *Melampsora larici-populina*

Antoine Persoons[1,2], Fabien Halkett[1,2] , Sébastien Duplessis[1,2] and Stéphane De Mita[1,2]

1 Institut National de la Recherche Agronomique, Unité Mixte de Recherche 1136 Institut National de la Recherche Agronomique/Université de Lorraine Interactions Arbres/Microorganismes, Champenoux, France

2 Université de Lorraine, Unité Mixte de Recherche 1136 Institut National de la Recherche Agronomique/Université de Lorraine Interactions Arbres/ Microorganismes, Vandoeuvre-lès-Nancy Cedex, France

Abstract :

Melampsora spp. are the most devastating pathogens for poplars worldwide, and, in particular, *Melampsora larici-populina* is a major threat for European poplar plantations (Pinon and Frey, 1997). In history, *M. larici-populina* has overcome almost all the qualitative resistances (denoted R1 to R8) released so far. The major resistance breakdown event occurred in 1994 and targeted the resistance R7. This event has led to the invasion of France by virulent 7 (i.e., able to successfully infect poplar cultivars exhibiting R7) *M. larici-populina* populations (Xhaard et al., 2011).

In this study we present a model of the demographic history of three genetic groups of *M. larici-populina* found in France: (1) a wild southern population, (2) a historical cultivated avirulent 7 cluster of populations, and (3) a virulent 7 cluster of populations which emerged more recently (Persoons et al., in prep.). We selected 86 isolates representing these three genetic groups and spanning the date of detection of R7 breakdown, and we detected genomic variation using Illumina-based whole-genome resequencing. We tested two scenarios modeling the ancestry of these three genetic groups using Approximate Bayesian Computation, representing alternative hypotheses concerning the origin of the virulent 7 population (from either the wild or the avirulent 7 populations) and estimated key demographic parameters. We found that the model connecting the virulent 7 cluster to the avirulent 7 rather than the wild population was more likely, with a significant divergence time between the two before the date of R7 breakdown.

We used a genome scan approach based on a combination of standard neutrality tests designed to detect specific regions of the genome where allelic frequencies have been skewed by the fast and recent fixation of a mutant (selective sweep) in the virulent 7 population specifically. This approach allowed the identification of 20 genomic regions showing potential signatures of selective sweeps. These regions contain 14 genes that could be considered as candidate genes related to the virulence 7.

INTRODUCTION

In most cases, evolutionary research focuses on historical events that occurred in a more or less distant past, such as domestication (Haudry et al., 2007; Li et al., 2014; Wang et al., 2014), divergence of populations (Punach and Stoneking, 2015; Begun et al., 2007; Liti et al., 2009) or adaptation to environmental conditions (Burke, 2012; Ellison et al., 2011; Namroud et al., 2008). Much less models allow to address the evolutionary process in a contemporary or at least very recent events. Due to their highly dynamic nature, host-pathogen interactions offer such opportunities. Indeed, host-pathogen interactions evolve constantly and rapidly, due to strong and reciprocal selection pressures, changing environmental conditions and, recently, anthropic interactions (Stukenbrock and Bataillon, 2012; Zaman et al., 2014).

Fungi of the Pucciniales order (Basidiomycota) cause rust diseases on a wide range of plants, including several economically important crop species (Duplessis et al., 2015). Among them, *Melampsora* spp. are the most devastating pathogens for poplars worldwide (Steenackers et al., 1996), and, in particular, *Melampsora larici-populina* is a major threat for European poplar plantations (Pinon and Frey, 1997). *Melampsora larici-populina* is a heteroecious rust which needs two hosts to complete its life cycle; *Populus* spp. on which it performs several asexual reproduction cycles each year during summer and autumn and *Larix decidua* on which it performs a single sexual reproduction cycle once a year in spring. *Melampsora larici-populina* is particularly damaging on cultivated cultivars of poplar mostly because of their intensive monoclonal cultivation over several decades (Gerard et al., 2006). In history, it has overcome almost all the qualtitative resistances (denoted R1 to R8) released so far. The major resistance breakdown event occurred in 1994 and targeted the resistance R7. The large impact of this breakdown event results from the massive planting of poplar cultivars carrying R7 in the 80's and 90's (in 1994, 80% of poplar planted in France carried R7; Pascal Frey, personnal communication). This event has led to the invasion of France by virulent 7 (i.e., able to successfully infect poplar cultivars exhibiting R7) *M. larici-populina* populations (Xhaard et al., 2011).

The population structure of *M. larici-populina* has been strongly impacted by R7 breakdown. A population genetic study based on genotyping and pathotyping of over 500 isolates sampled in France from 1992 to 2011 allowed the identification of three genetics groups, of which two were impacted by the R7 breakdown (Persoons et al., in prep.). A group is found in the South of France. Geographically, isolates falling into this group were sampled within the area of sympatry of the wild poplar *Populus nigra* (telial host) and of *Larix decidua* (aecial host), with large populations of both hosts. In contrast, cultivated poplar stands (in particular those harboring R7) are scarce, suggesting that the effect of R7 breakdown on members of this wild genetic group was probably

minor. Nearly all isolates falling into this wild group have been found to be avirulent 7, i.e., unable to infect R7 poplar cultivars. A second group is present in North-East of France and has disappeared in 1997 (within the limits of our sample). This group comprises avirulent 7 isolates only, although they bear other virulences (i.e., they can infect poplar cultivars bearing formerly deployed resistances). The last group comprises virulent 7 isolates and is not detected before 1994. It presents genetic signatures of a bottleneck followed by demographic expansion (Persoons et al., in prep.). Taken together, these results indicate that the R7 breakdown had a strong demographic consequence on *M. larici-populina* populations, and suggest a scenario including the replacement of the ancestral avirulent 7 genetic group by the modern virulent 7 group, at least in the regions where poplar is cultivated (which represents, in France, most of the territory). The sampling in the study reported in Persoons et al., (in prep.) might not have been extensive enough, both in broadness and density, to uncover remaining avirulent 7 populations in the poplar cultivation area, but it seems likely that the virulent 7 group has taken over to a wide extent, in most of France.

Obviously, the driving force responsible for the hypothesized replacement of the ancestral avirulent 7 population by the virulent 7 population is natural selection. Indeed, a though the ecosystem and the conditions in that case are shaped by humans, the process of adaptation is essentially unguided and therefore natural. Most probably, R7 overcoming has been mediated by changes at the genetic level. It is classically assumed that the molecular evolution of host-pathogen interactions follow a gene-for-gene model where two single genetic loci, one in each partner, interact with each other in such a way that mutations in either one can prevent infection (from the host's point of view) or secure it (from the pathogen's point of view) (Flor, 1971). Recurring coevolution will lead to repeated changes at the two loci. The molecular bases of interspecific interactions are likely more complex than this simple scenario, and changes at other loci can also counteract the adaptations of the other interacting partner, such that, in the end, several loci can interact in cascade potentially increasing the opportunities for evolution (Jones and Dangl, 2006). Nevertheless, a single resistance breakdown will necessarily reduce to a gene-for-gene instance of evolution, even within a complex cascade of interactive molecular partners. However, nothing demonstrates that the R7 breakdown was mediated by a single genetic change or, for that matter, that R7 itself is determined by a single locus.

Identifying genomic regions subjected to selection has been a longstanding interest of evolutionary biologists (Lewontin and Krakuer, 1973). The recent emergence of high-throughput sequencing technologies offering a wide genomic coverage (Mardis, 2008) provides opportunities to detect loci involved in adaptation (Ellegren, 2013). Modern sequencing techniques allowed both to dramatically increase the number of loci characterized and to progressively increase the number of individuals used in population samples. Many methods have been designed to detect the

signatures of adaptation, long before genome sequencing became routine (Nielsen, 2005) and, recent methodologies have been introduced to leverage the wealth of data that became available thanks to whole-genome sequencing (Vitti et al., 2013).

The evolutionary event represented by R7 breakdown in French *M. larici-populina* population has both demographic and adaptive implications. As for many pathogen species, *M. larici-populina* has a partially clonal life cycle that enables rapid population growth and colonization of new environments (MacDonald and Linde, 2002). This clonal phase of the life cycle has potentially played a role in the speed of the spread of virulent 7 genotypes. However, a sexual reproduction cycle is imposed annually which, together with the obligatory host change, is likely to enhance gene flow and promote recombination (Xhaard et al., 2011).

In this study we present a model of the demographic history of the three genetic groups of *M. larici-populina* found in France, namely the wild Southern population, the historical cultivated avirulent 7 cluster, and the more recently identified virulent 7 cluster. Our study makes use of the historical collection of *M. larici-populina* available for multiplication and genomic sequencing. We selected 86 isolates representing the three genetic groups and spanning the date of detection of R7 breakdown and detected genomic variation using Illumina-based whole-genome resequencing (yielding over a million single-nucleotide polymorphisms [SNPs]). We tested two scenarios modeling the ancestry of these three genetic groups using Approximate Bayesian computation (ABC; Beaumont et al., 2002). The analysis is based on simulations using an algorithm that makes two questionable assumptions: the absence of selection and panmictic reproduction. However, we assume here that the annual cycle of sexual reproduction is sufficient to make the natural populations behaves as panmictic units, as suggested both theoretically (Balloux et al., 2003; Halkett et al., 2005; De Meeus et al., 2007; Stoeckel and Masson, 2012) and by experimental data gained on this biological model (Barrès et al., 2008 , 2012; Xhaard et al., 2011, and Persoons et al., in prep.), and that recombination is sufficient to unphase most of the genome from the locus (or the loci) involved R7 breakdown (Persoons et al., in prep.). Using the modeling setup, we tested alternative hypotheses of the origin of the virulent 7 population (from either the wild or the avirulent 7 population) and estimated key parameters: effective population size, strength of the bottleneck caused by R7 breakdown, time of divergence of virulent 7 population, and rate of gene flow between populations.

The setting of the R7 breakdown offers a favorable situation to identify the molecular determinants of the adaptation to the host, in particular thanks to the recent event of adaptation, the availability of samples predating and immediately following the event, and overall the presence of genetic recombination making the detection of individual loci possible. We used a genome scan approach based on a combination of standard neutrality tests designed to detect specific regions of

the genomes where allelic frequencies have been skewed by the fast and recent fixation of a mutant (selective sweep) in the virulent 7 population specifically. The expected distribution of test statistics has been generated using simulations under the most likely model and integrating the posterior probability of parameters, based on genomic windows of fixed sizes.

MATERIALS AND METHODS

Fungal material

Isolates were selected from a laboratory collection (Frey P., INRA Nancy, Champenoux, France) representing more than 20 years of *Melampsora larici-populina* strains sampled from poplar trees (both wild *Populus nigra* and *P.* spp cultivars) throughout France. Four populations have been defined based on the structure previously described in Persoons et al., (in prep) (table 1).

The four populations used in this study are: (1) a sampling of 21 isolates from the avirulent 7 group sampled in eastern France (where cultivated poplar is frequent) in 1993 (*Avr7* population), (2) a sampling of 21 isolates from the virulent 7 group sampled in eastern and northern France in 1994 (*94Vir7* population), (3) a sampling of 22 isolates from the virulent 7 group sampled in eastern and northern France in 1998 (*98Vir7* population), and (4) a sampling of 22 isolates from the wild group sampled in South-Eastern France in 2008 (*Wild* population).

In order to avoid contamination, one urediniospore of each isolate was multiplied on detached leaves of trees of the Robusta cultivar. Pathotypes of all isolates (i.e., combination of virulences) were confirmed in triplicate experiments on eight poplar cultivars each carrying a single resistance (R1 to R8) to *M. larici-populina* (table 1) and on the sensible Robusta cultivar, as a positive control. To ensure their purity and to avoid potential clones within the selected isolates, genotyping was performed using 25 microsatellite markers (Xhaard et al., 2009).

DNA isolation

A total of 100-300 mg of urediniospores were used for DNA isolation using the CTAB method. Spores were crushed using a Retsch Tissue Lyser (Qiagen, Courtaboeuf, France) at a frequency of 30 Hz during 1 min. Spores were resuspended in CTAB buffer (Tris 0.1 M, NaCl 1.43 M, EDTA 0.02 M, CTAB 0.02 M) and heated at 65°C for 30 min. The suspension was subjected to centrifugation at 8000 rpm at room temperature for 5 min to pellet spore debris. Supernatant was gently mixed with an equal volume of phenol:chloroform:isoamyl alcohol (50:48:2; Euromedex,Souffelweyersheim, France) and centrifuged at 8000 rpm at room temperature for 10 min. The aqueous phase was recovered, gently mixed with an equal volume of chloroform and

Sample name	Location	Sampling year	Pathotype	Genetic group	Sequencing	Tot read	Mapped read	Unmapped read	Sequencing_Depth
93CU1	Amance, Lorraine, France	1993	2-4	Avr7	JGI	142 725 766	106 234 916	36 490 850	158
93CV1	Amance, Lorraine, France	1993	1-3-4-5	Avr7	JGI	65 954 482	48 475 540	17 478 942	72
93DL2	Amance, Lorraine, France	1993	2-3-4-5	Avr7	JGI	65 011 844	31 035 862	33 975 982	46
93EA2	Amance, Lorraine, France	1993	3-4-6	Avr7	JGI	101 349 334	83 170 674	18 178 660	124
93EC4	Amance, Lorraine, France	1993	3-4	Avr7	Beckman	59 910 186	52 768 585	7 141 601	78
93EF8	Amance, Lorraine, France	1993	1-3-4-6	Avr7	JGI	65 183 830	55 521 367	9 662 463	82
93GR7	Amance, Lorraine, France	1993	3-4-6	Avr7	JGI	72 757 404	53 748 378	19 009 026	80
93GS3	Amance, Lorraine, France	1993	2-4	Avr7	Beckman	86 362 164	76 270 750	10 091 414	113
93GT1	Amance, Lorraine, France	1993	2-4	Avr7	JGI	72 556 526	59 368 999	13 187 527	88
93GV8	Amance, Lorraine, France	1993	2-4	Avr7	JGI	75 063 724	49 024 774	26 038 950	73
93GW3	Amance, Lorraine, France	1993	2-4	Avr7	Beckman	53 957 290	46 027 158	7 930 132	68
93HQ8	Amance, Lorraine, France	1993	2-4	Avr7	JGI	59 303 026	46 033 052	13 269 974	68
93HW5	Amance, Lorraine, France	1993	1-3-4-5	Avr7	Beckman	57 892 188	48 970 094	8 922 094	73
93HY5	Amance, Lorraine, France	1993	3-4	Avr7	Beckman	37 814 402	32 136 124	5 678 278	48
93IB7	Amance, Lorraine, France	1993	1-3-4-5	Avr7	JGI	65 013 360	46 692 911	18 320 449	69
93IC8	Amance, Lorraine, France	1993	2-4	Avr7	JGI	70 075 734	54 341 831	15 733 903	81
93NU2	Amance, Lorraine, France	1993	3-4	Avr7	JGI	74 021 678	58 078 328	15 943 350	86
93NU4	Amance, Lorraine, France	1993	1-3-4-5	Avr7	JGI	68 789 184	54 282 443	14 506 741	81
93NX6	Amance, Lorraine, France	1993	2-3-4-5	Avr7	JGI	71 550 396	51 128 752	20 421 644	76
93OC5	Amance, Lorraine, France	1993	2-4	Avr7	Beckman	56 235 566	47 581 671	8 653 895	71
93Q1	Amance, Lorraine, France	1993	2-3-4	Avr7	Beckman	94 027 938	56 451 893	37 576 045	84
94zz1	ETAIN, Lorraine, France	1994	1-3-4-5-7	Vir7	Beckman	84 580 140	70 730 268	13 849 872	105
94ZZ10	Grammont, Belgique	1994	1-3-4-5-7	Vir7	JGI	57 148 512	49 488 033	7 660 479	73
94ZZ11	ONNAING, Nord-Pas-de-calais, France	1994	1-3-4-5-7	Vir7	JGI	67 034 324	53 181 210	13 853 114	79
94ZZ12	LECELLES, Nord-Pas-de-calais, France	1994	1-3-4-5-7	Vir7	Beckman	48 363 596	41 283 024	7 080 572	61
94ZZ13	SAULCHOY, Nord-Pas-de-calais, France	1994	3-4-5-7	Vir7	JGI	98 606 918	82 756 190	15 850 728	123
94ZZ14	Grammont, Belgique	1994	1-3-4-5-7	Vir7	JGI	80 351 982	57 794 157	22 557 825	86
94ZZ15	Hesdin, Nord-pas-de-Calais, France	1994	3-5-7	Vir7	JGI	56 099 432	48 553 209	7 546 223	72
94ZZ16	LECELLES, Nord-Pas-de-calais, France	1994	1-3-4-5-7	Vir7	JGI	75 102 408	61 086 729	14 015 679	91
94ZZ17	SAULCHOY, Nord-Pas-de-calais, France	1994	1-3-4-5-7	Vir7	JGI	70 927 994	54 561 820	16 366 174	81
94ZZ18	ETAIN, Lorraine, France	1994	1-3-4-5-7	Vir7	JGI	117 150 160	85 335 455	31 814 705	127
94zz19	Grammont, Belgique	1994	1-3-4-5-7	Vir7	Beckman	50 336 594	42 017 309	8 319 285	62
94zz2	ONNAING, Nord-Pas-de-calais, France	1994	3-4-5-7	Vir7	Beckman	54 919 794	44 887 073	10 032 721	67
94ZZ20	Guesnain, Nord-Pas-de-calais, France	1994	1-3-4-5-7	Vir7	JGI	71 836 564	63 888 549	7 948 015	95
94zz21	SAULCHOY, Nord-Pas-de-calais, France	1994	1-3-4-5-7	Vir7	Beckman	26 555 924	19 785 181	6 770 743	29
94ZZ28	Nogent, Centre, France	1994	1-3-4-5-7	Vir7	JGI	66 639 506	55 447 361	11 192 145	82
94ZZ3	Amance, Lorraine, France	1994	3-4-5-7	Vir7	JGI	73 726 742	55 996 026	17 730 716	83
94ZZ33	ETAIN, Lorraine, France	1994	1-3-4-5-7	Vir7	JGI	74 346 320	55 042 524	19 303 796	82
94ZZ34	Grammont, Belgique	1994	1-3-4-5-7	Vir7	JGI	70 110 142	58 197 974	11 912 168	86
94ZZ35	Grammont, Belgique	1994	1-3-4-5-7	Vir7	JGI	68 656 552	51 526 704	17 129 848	77
94ZZ36	Grammont, Belgique	1994	1-3-4-5-7	Vir7	JGI	74 073 542	58 392 690	15 680 852	87
94ZZ37	Grammont, Belgique	1994	1-3-4-5-7	Vir7	JGI	69 019 126	49 899 404	19 119 722	74
94zz7	ETAIN, Lorraine, France	1994	1-3-4-5-7	Vir7	Beckman	43 217 514	36 324 477	6 893 037	54
98AA6	Amance, Lorraine, France	1998	1-3-4-5-7	Vir7	JGI	67 949 182	57 950 943	9 998 239	86
98AB02	Amance, Lorraine, France	1998	1-4-5-7	Vir7	Beckman	30 270 568	26 854 716	3 415 852	40
98AB07	Amance, Lorraine, France	1998	1-3-4-5	Vir7	Beckman	68 748 766	58 669 384	10 079 382	87
98AC08	Amance, Lorraine, France	1998	1-4-5-7	Vir7	JGI	52 627 272	36 954 213	15 673 059	55
98AD04	Amance, Lorraine, France	1998	1-4-5-7	Vir7	Beckman	69 531 658	59 280 231	10 251 427	88
98AG18	Moy-de-l'Aisne, Picardie, France	1998	1-3-4-5-7	Vir7	Beckman	87 354 362	75 004 501	12 349 861	111
98AG3	Moy-de-l'Aisne, Picardie, France	1998	1-3-4-5-7	Vir7	JGI	71 452 586	57 264 444	14 188 142	85
98AG36	Moy-de-l'Aisne, Picardie, France	1998	1-3-4-5-7	Vir7	Beckman	30 239 348	26 065 324	4 174 024	39
98AG42	Moy-de-l'Aisne, Picardie, France	1998	3-4-7	Vir7	JGI	74 168 356	61 762 059	12 406 297	92
98AG49	Moy-de-l'Aisne, Picardie, France	1998	1-3-4-5-7	Vir7	JGI	64 249 402	47 623 027	16 626 375	71
98AG54	Moy-de-l'Aisne, Picardie, France	1998	1-3-4-5-7	Vir7	JGI	60 253 650	53 397 678	6 855 972	79
98AG56	Moy-de-l'Aisne, Picardie, France	1998	3-4-7	Vir7	JGI	74 682 036	52 160 759	22 521 277	77
98AG69	Moy-de-l'Aisne, Picardie, France	1998	1-3-4-5-7	Vir7	Beckman	25 737 736	21 327 218	4 410 518	32
98AI1	Moy-de-l'Aisne, Picardie, France	1998	1-3-4-5-7	Vir7	JGI	69 824 760	54 788 139	15 036 621	81
98AI35	Moy-de-l'Aisne, Picardie, France	1998	3-4-7	Vir7	JGI	79 127 386	64 085 440	15 041 946	95
98AI58	Moy-de-l'Aisne, Picardie, France	1998	1-3-4-5-7	Vir7	JGI	77 007 060	64 983 378	12 023 682	97
98AP03	Amance, Lorraine, France	1998	1-4-5-7	Vir7	Beckman	62 538 212	45 704 534	16 833 678	68
98AP2	Amance, Lorraine, France	1998	1-3-4-5-7	Vir7	JGI	78 813 448	69 178 565	9 634 883	103
98GB3	Amance, Lorraine, France	1998	1-3-4-5-7	Vir7	JGI	47 005 584	33 874 603	13 130 981	50
98GC04	Amance, Lorraine, France	1998	1-5-6-7	Vir7	Beckman	31 451 444	26 308 970	5 142 474	39
98GD5	Amance, Lorraine, France	1998	3-4-7	Vir7	JGI	58 574 292	45 039 555	13 534 737	67
98GF7	Amance, Lorraine, France	1998	1-4-5-7	Vir7	JGI	67 013 632	56 269 608	10 744 024	84

08EA01	Prelles, Provence-alpes côte d'azur, France	2008	4	Wild	JGI	74 819 586	58 099 506	16 720 080	86
08EA08	Prelles, Provence-alpes côte d'azur, France	2008	4	Wild	JGI	73 381 290	60 591 216	12 790 074	90
08EA09	Prelles, Provence-alpes côte d'azur, France	2008	0	Wild	JGI	74 676 036	55 185 256	19 490 780	82
08EA105	Prelles, Provence-alpes côte d'azur, France	2008	4	Wild	Beckman	31 320 890	27 400 363	3 920 527	41
08EA106	Prelles, Provence-alpes côte d'azur, France	2008	4	Wild	JGI	76 034 268	51 138 349	24 895 919	76
08EA11	Prelles, Provence-alpes côte d'azur, France	2008	4	Wild	JGI	72 144 222	48 933 250	23 210 972	73
08EA15	Prelles, Provence-alpes côte d'azur, France	2008	2-4	Wild	JGI	74 342 762	58 354 923	15 987 839	87
08EA16	Prelles, Provence-alpes côte d'azur, France	2008	4	Wild	JGI	60 232 364	33 854 700	26 377 664	50
08EA18	Prelles, Provence-alpes côte d'azur, France	2008	0	Wild	JGI	67 312 896	56 577 920	10 734 976	84
08EA19	Prelles, Provence-alpes côte d'azur, France	2008	4	Wild	JGI	113 044 060	98 034 745	15 009 315	146
08EA22	Prelles, Provence-alpes côte d'azur, France	2008	4	Wild	JGI	73 424 058	51 752 625	21 671 433	77
08EA31	Prelles, Provence-alpes côte d'azur, France	2008	4	Wild	JGI	66 002 342	38 437 462	27 564 880	57
08EA42	Prelles, Provence-alpes côte d'azur, France	2008	4	Wild	JGI	63 165 170	50 633 512	12 531 658	75
08EA43	Prelles, Provence-alpes côte d'azur, France	2008	0	Wild	JGI	75 944 106	65 002 401	10 941 705	97
08EA50	Prelles, Provence-alpes côte d'azur, France	2008	2-4	Wild	JGI	62 437 714	53 725 471	8 712 243	80
08EA55	Prelles, Provence-alpes côte d'azur, France	2008	4	Wild	JGI	70 595 156	56 811 312	13 783 844	84
08EA6	Prelles, Provence-alpes côte d'azur, France	2008	4	Wild	JGI	104 526 526	89 222 727	15 303 799	133
08EA61	Prelles, Provence-alpes côte d'azur, France	2008	0	Wild	JGI	67 448 882	52 677 496	14 771 386	78
08EA66	Prelles, Provence-alpes côte d'azur, France	2008	4	Wild	Beckman	65 783 510	51 130 282	14 653 228	76
08EA85	Prelles, Provence-alpes côte d'azur, France	2008	4	Wild	JGI	58 689 976	30 785 174	27 904 802	46
08EA95	Prelles, Provence-alpes côte d'azur, France	2008	0	Wild	Beckman	54 997 764	44 242 083	10 755 681	66

Table 1: Summary of *Melampsora larici-populina* isolates. Isolate name, year and location of sampling are indicated. The pathotype profile (combination of virulences) was confirmed in triplicate by inoculation on a differential set of poplar cultivars carrying the eight known resistances to *M. Larici-populina*. The genome of these isolates was sequenced by two companies (sequencing) with the same 2*150pb Illumina paired end method. The delivery reads (tot reads) were mapped on the reference genome. The number of mapped and unmapped reads are indicated with the corresponding sequencing depth.

centrifuged at 8000 rpm at room temperature for 10 min. The aqueous phase was subjected to RNA digestion with RNaseA at 10 µM (Fermentas, Saint-Remy-lès-Chevreuse, France) in excess at 37°C for 30 min. A final extraction with an equal volume of chloroform was realized followed by centrifugation at 8000 rpm at room temperature for 10 min. The aqueous phase recovered was then subjected to isopropanol (0.75 of final volume) precipitation, followed by centrifugation at 14 000 rpm at 4°C during 30 min. DNA pellet was washed twice with 70%, then absolute ethanol, each followed by centrifugation at 14 000 rpm at 4°C for 10 min. The DNA pellet was finally dried under a hood for 20 min and resuspended in 1X Tris EDTA. Quality and quantity of recovered high molecular weight DNA was assessed by electrophoresis on agarose gel, by spectrophotometry (Nanodrop, Saint-Remy-lès-Chevreuse, France) and with the QuBit (Life Technologies, Villebon-sur-Yvette, France) fluorometric quantitation system.

Genome re-sequencing, filtering and mapping of short reads

Genomic DNA libraries were used for sequencing by Beckman Coulter Genomics (Grenoble, France) for 24 samples and by the Joint Genome Institute (Walnut Creek, California, United States) for others (table 1). Each library was quantified by qPCR and sequenced on the Illumina HiSeq2000 platform as paired-end 150 bp reads. The total number of sequencing reads for each sample is presented in table 1.

For the mapping we used the new version of the *M. larici-populina* reference genome generated by the Joint Genome Institute (Jeremy Schmutz and Jerry Jenkins, pers. comm.). This genome is assembled in 18 linkage groups (versus 462 scaffolds for the current released version) based on the genetic map obtained from the reference isolate auto-cross (Pernaci et al., 2014). The length of each linkage group is presented in table 2.

The forward and reverse libraries of each sample were aligned on the reference genome using the bwa software version 0.7.5 (Li and Durbin, 2009). We fixed the maximum number of differences for each read against the reference to 2. All others options were used with default values. The number of mapped and unmapped reads and the sequencing depth by isolate were computed with samtools version 0.1.19 (Li et al., 2009) (table 1). Variant call format files (vcf) were generated for each linkage group using samtools with default parameters. SNP calling was performed using a custom filtering procedure implemented in Python using EggLib version 3.0.0 alpha (De Mita and Siol, 2012). Filter parameters were: minimum minor allelic frequency = 0.05; minimum sequencing depth = 15; maximum genotype likelihood for each sample: 20.

Analysis of polymorphism data

Linkage disequilibrium (LD) was computed for all pairs of SNPs within a given linkage group as

Linkage group	Length	SNP	INDEL	Total
LG01	8 567 145	83 617	4 078	87 695
LG02	7 400 714	87 500	4 068	91 568
LG03	7 995 696	91 839	4 572	96 411
LG04	6 413 982	64 814	3 103	67 917
LG05	6 145 269	59 158	2 952	62 110
LG06	7 015 942	74 430	3 607	78 037
LG07	5 819 754	60 697	3 104	63 801
LG08	6 014 601	75 302	3 800	79 102
LG09	4 829 805	51 796	2 681	54 477
LG10	4 781 055	55 621	3 038	58 659
LG11	5 560 581	50 867	2 359	53 226
LG12	4 992 572	55 051	2 611	57 662
LG13	4 013 672	47 828	2 365	50 193
LG14	4 116 462	42 991	2 060	45 051
LG15	4 671 214	52 499	2 404	54 903
LG16	4 243 487	52 101	2 257	54 358
LG17	4 451 399	53 387	2 383	55 770
LG18	4 402 489	50 148	2 367	52 515
Total	101 435 839	1 109 646	53 809	1 163 455

Table 2: Characteristics of the mapping of reads on the 18 linkage groups of the new *M. larici-populina* genome. The number of Single Nucleotide Polymorphisms (SNP) and Insertion/deletion (INDEL) detected for the 86 isolates against each linkage group are indicated.

the squared correlation coefficient. Values were obtained for the whole data set and for each population separately. Next, we computed genetic diversity statistics in windows of 5 kb in keeping only bi-allelic SNP with at least 1 SNP. The number of windows was 17 805 with a maximal number of SNPs per window of 180 and an average of 29.7 (total: 528 517 SNPs). Statistics were: Tajima's *D* (Tajima, 1989), Watterson's θ estimator (thetaW) (Watterson, 1975) and F_{ST} (Weir and Cockerham, 1984). We also computed Tajima's *D* in the virulent 7 group (grouping *94Vir7* and *98Vir7* samples; *D*vir) and in the avirulent 7 group (grouping *Wild* and *Avr7* samples; *D*avr), and F_{ST} when only considering the partition of the virulent 7 group against the avirulent 7 group (later denoted as F_{ST} Vir-Avr).

With the aim of performing demographic inference based on regions representing the whole genome with minimal physical linkage between them, we selected 476 windows of 2 kb containing at least 1 bi-allelic SNP and separated by at least 200 kb. The number of bi-allelic SNPs per window ranged from 1 to 243 with an average of 21.0 (total 10 014 SNPs). Based on SNPs found in these windows, we computed the folded site frequency spectrum (SFS). The total sample size was 86 individuals, each bearing two copies of the genome (isolates were at a dicaryotic stage of their life cycle at the time of DNA extraction, which is not different than a diploid after DNA extraction; from now on we will use the term diploid for simplicy). Therefore the maximal MAF was 86. The minimal MAF was 9 due to the frequency filter of 0.05. To reduce sampling variance, we used 16 categories of allelic frequencies: 9-13;14-18;19-36; 24-28; 29-33; 34-38; 39-43; 44-48; 49-53; 54-58; 59-63; 64-68; 69-73; 74-78; 79-83; 84-86. In addition, we computed joint frequency spectrum (JFS) for three pairwise comparisons: *94Vir7* against *98Vir7*; *Wild* against *94Vir7* and *Wild* against *Avr7*. All JFSs were computed over a 6×6 grid using the following MAF categories: 9-21; 22-34; 35-47; 48-60; 61-73; 74-86. All statistics have been computed using EggLib version 3.0.0 alpha.

ABC analysis

We designed two demographic scenarios described by 7 parameters each using the coalescent simulator included in EggLib. Demographic scenarios and parameters are presented in figure 5. Each of the scenarios features two population splits (the difference of the two scenarios consists in order), a population bottleneck in the virulent 7 population, permanent migration rate between populations and four samples (including three past samples). The free parameters are *N*, the size of the wild population (used as reference), expressed in diploid individuals, the size of both virulent 7 (*N1*) and avirulent 7 (*N2*) populations (both expressed relatively to *N*), the migration rate *M* (equal to *4Nm* where *m* is the per-generation migration rate), the bottleneck strength *S* (expressed as the number of generations, times 4*N*, necessary to obtain the equivalent number of coalescence events)

and two waiting times before both split events (*T1* and *T2*), expressed in numbers of generations times *4N*. The migration matrix is decomposed as followed: between the virulent 7 and avirulent 7 populations: fixed to *M* (symmetrical), between the wild and the virulent 7 or avirulent 7 populations: fixed to *M*/100 (also symmetrical). The waiting time to all sampling dates was fixed to the known values. Time is expressed in 4*N* generations and we assume one (sexual) generation per year. The present time (most recent sampling) is 2008, so the waiting times are 10 years (2.5/*N*) for *98Vir7*, 14 years (3.5/*N*) for *94Vir7* and 15 years (3.75/*N*) for *Avr7*. The bottleneck date is supposed to be more ancient than the *98Vir7* sampling but the time between the two event is supposed to be small enough to be neglected. The prior distributions of free parameters are presented in table 3 and appear in figure 7 (blue lines).

We drawn approximately 1.5×10^{6} parameter values from the prior for each of two models (1 435 000 and 1 688 000, respectively). For each parameter set, we simulated an alignment for every of the 476 windows, each being conditioned to the corresponding number of SNPs in an infinite-site model of mutation. The number of mutations has therefore been fixed and no parameter has been implemented to represent the mutation rate. All windows have been assumed to be independent and have been simulated separately. The recombination rate windows was fixed to 0. For all simulated dataset, we computed the SFS and the three JFSs as described for the empirical data.

ABC analyses were performed with artificial neural networks as available in the ABC R package version 2.1 (Csilléry et al., 2012). For model comparison, 16 different tolerance values ranging from 5×10^{-4} to 8×10^{-3} and with default options values expected for a larger maximum number of iterations (5000). Different subsets of simulated values were used (250 000, 500 000 and 750 000 simulations per model). Model probabilities and Bayes factors were retrieved from the ABC package outcome. For estimating parameters, we used a tolerance of 1.5×10^{-3} (more than 2000 simulations in the accepted area for both models), logit transformation and otherwise default values of parameters.

Posterior simulations and genome scan

The marginal distribution of each parameter was discretized using a variable number of categories (*N*: 15; *T1*: 50; *T2*: 40; *N1*: 15; *N2*: 15; *S*: 15; *M*: 20). The number of categories was higher for parameters which had a sharper mode. To perform posterior simulations (simulations assuming one of the model, and taking into account the uncertainty over parameter values), we drawn randomly categories from the discretized posterior distributions and, within categories, we draw values using uniform distributions. We performed simulations conditioned on the number of SNPs found in windows (1 000 000 replicates for each value, ranging from 1 to 180 SNPs per windows). For all

simulations, the following statistics were computed: *D*, F_{ST}, *D*vir, *D*avr, F_{ST} Vir-Avr and thetaW. The distribution of statistics obtained from these posterior simulations were used to perform the statistical test of neutrality for all 5 kb windows covering the whole genome. We specifically selected candidate windows for which both *D*vir and F_{ST} Vir-Avr were significant. Finally, we extracted genes falling into candidate windows based on the gene models of the current version of the *M. larici-populina* genome that have been found unmodified in the most recent version (11 604 genes found out of 16 399).

RESULTS

Genome re-sequencing and SNP calling

The genome of 86 *M. larici-populina* isolates was sequenced with the Illumina technology. An average of 68 million reads were generated, corresponding to 25 to 143 million reads per isolate. The number of allowed differences used for mapping was set to a stringent value (two differences). On average, 78% of the reads aligned with the 18 linkage groups of the reference genome. Overall, this led to a sequencing depth average of 79X per isolate (32X to 158X; table 1). The SNP detection procedure identified a total of 1 163 455 variants including 1 109 646 SNPs and 53 809 insertions/deletions (table 2). In order to avoid false positives caused by sequencing errors at high depth, we conserved SNPs without missing data, with only two alleles and with minority allele frequency superior to 0.05, which dropped a little bit more than half of SNPs (final number of 528 517).

Analysis of polymorphism

We defined 10 014 windows along the 18 linkage groups and computed Tajima's *D*, F_{ST} , *D*vir (grouping *94Vir7* and *98Vir7* samples), *D*avr (grouping *Wild* and *Avr7* samples) and F_{ST} Vir-Avr (virulent 7 group against avirulent 7 group) (figure 1). Interestingly, Tajima's *D* distributions are centered on positive values for the global dataset and for both *D*vir and *D*avr, which can be explained by both population sub-structure and the strong frequency threshold that has been applied. The variance of *D* values is wide, and is inflated towards positive values for the global dataset and on both directions for both *D*vir and *D*avr. The mode of the global F_{ST} distribution is below 0.1 and F_{ST} Vir-Avr less than 0.05, although in both cases values over 0.2 occur at appreciable frequencies.

We then computed linkage disequilibrium using all data and within each sample (figure 2). We observed that LD decays over essentially 200 Kb although some pairs of SNPs with high linkage are still found in the global dataset. However there is a marked difference in the rate of LD

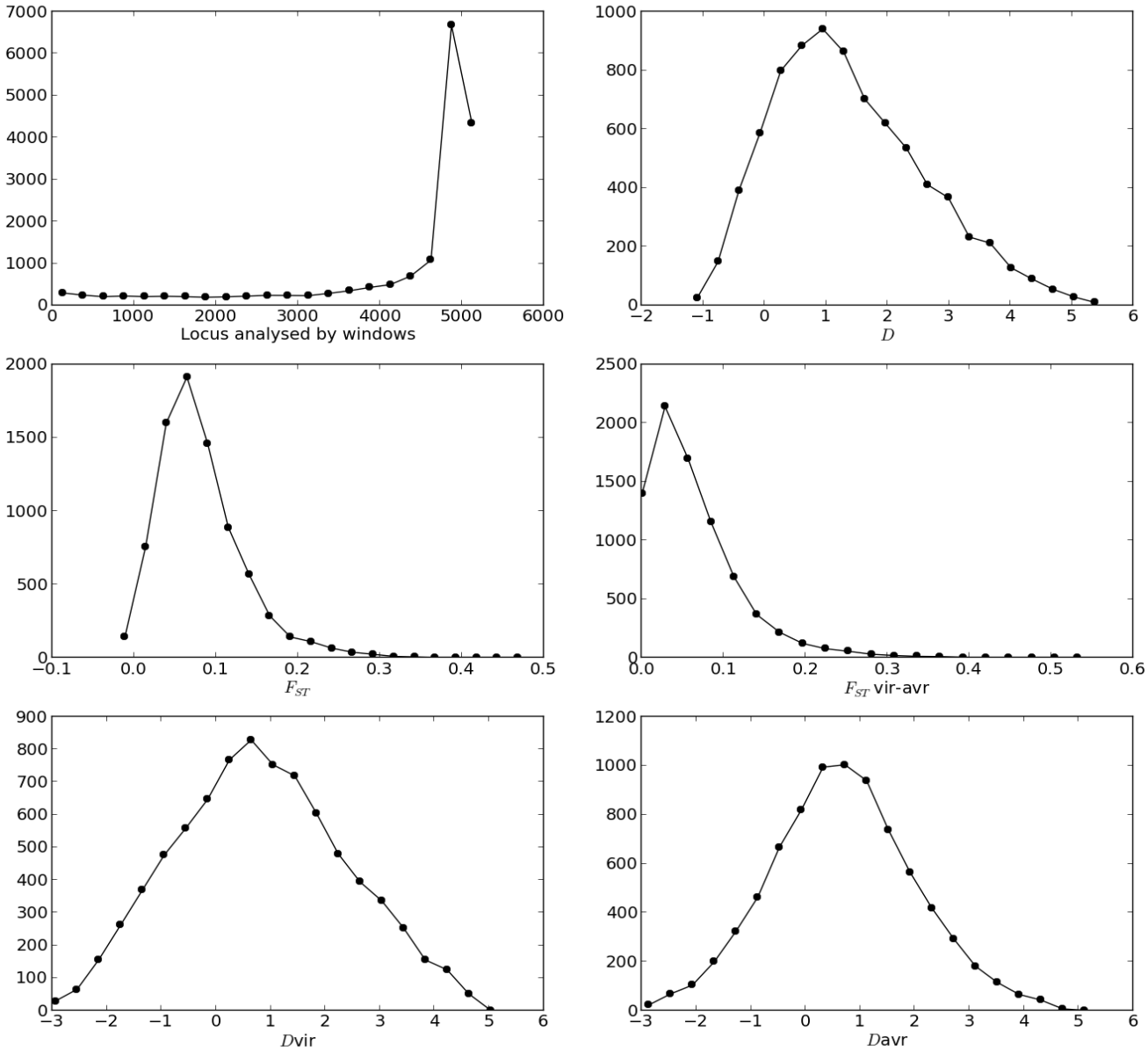


Figure 1: Statistical analyse of 5kb windows with number of locus analysed (A), Tajima's *D* (B), F_{ST} (C), F_{ST} Vir-Avr (D), *D*vir (E) and *D*avr (F).

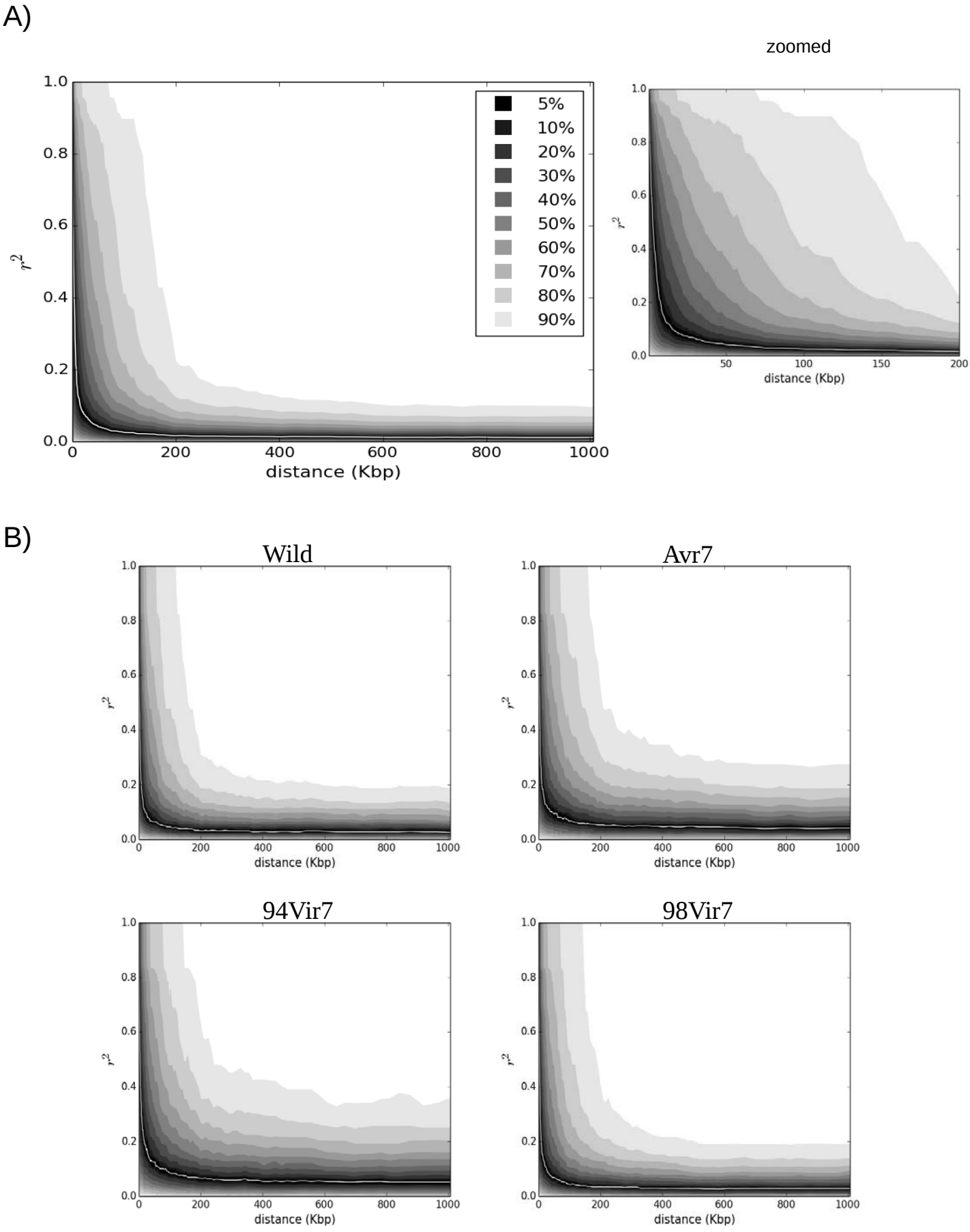


Figure 2: Linkage desequilibirum computed for all SNPs pairs on all isolates (A) and by populations (B). The grayscale is the amount of SNPs pairs present in each data range (corresponding to the scale of Figure A). The white line represents the average data In the four bottom panels, small random noise has been added to r^2 values in order to prevent point superposition.

decay between the different samples, with slower decay for the *Avr7* and especially the *94Vir7* samples. The *Wild* group presents comparatively less long-distance genetic linkage.

Based on the result obtained for LD, we selected 476 windows of 2 kb separated by at least 200 kb, with the aim of generating a dataset easy to reproduce in simulations. Based on these windows, we generated the SFS, as presented in figure 3. As expected, much of the distribution is represented by lower frequencies, although frequencies lower than 9 gene copies have not be examined here. A unexplained peak was observed for the category gathering alleles at 54 up to 58 copies. This class of frequencies does not correspond to any between-population divergence, and it has no effect on the 16-class SFS used in the ABC analysis below.

The JFS was computed for *94Vir7* against *Wild*, *98Vir7* against *94Vir7* and *Avr7* against *Wild* sample comparisons. Each JFS was categorized in 6×6 categories (total 36 categories) in order to reduce noise in the signal for the ABC analysis (figure 4). There is a stronger diagonal for the JFS of *94Vir7* against *98Vir7*, indicating a stronger correlation of allele frequencies which is expected because these samples come from the same genetic group.

ABC analysis

We designed two demographic scenarios (A and B) with 7 parameters (table 3) and described in figure 5. We performed model choice comparing to observed data (shown in figure 6 and figure 7). Three subsets of data have been used (250 000, 500 000 and 750 000 simulations per scenario) and 16 different tolerance values (5×10^{-4} to 8×10^{-3}) have been used (figure 6). For tolerance values below 2×10^{-3}, model A is unambiguously supported but, for higher (less stringent) tolerance values, the two models are less easy to discriminate. Counter-intuitively, more power to discriminate is achieved with less data (even for a given tolerance level), suggesting that large amounts of data cause computational problems.

Based on the result of model choice, we choose the model A and computed the posterior distribution of parameters under this model (figure 7). The majority of parameters present a much narrow (sometimes shifted) posterior distribution than the prior, indicating that signal was available to estimate those parameters. For *N* (number of individuals in the wild population), the distribution is not much shifted compared with the prior (average 6445 compared with 6471 for the prior), but the mode is much sharper (95% credible interval [CI]: 2230-10 926). Both waiting times are estimated with a much larger precision than the prior (*T1*, time between the bottleneck and the first population split, CI: 0.0402-0.1291, average: 0.0756; *T2*, time between the two population splits, CI: 0.00278-0.00763, average: 0.00481), as well as *S* (the strength of the bottleneck) with a CI of 0.0667-0.1034 and an average of 0.0853 and, to a lesser extent, *M* (the migration rate) with a CI of 0.215-1.206 and an average of 0.697. In comparison, the avirulent 7 (*N1*) and virulent 7 (*N2*)

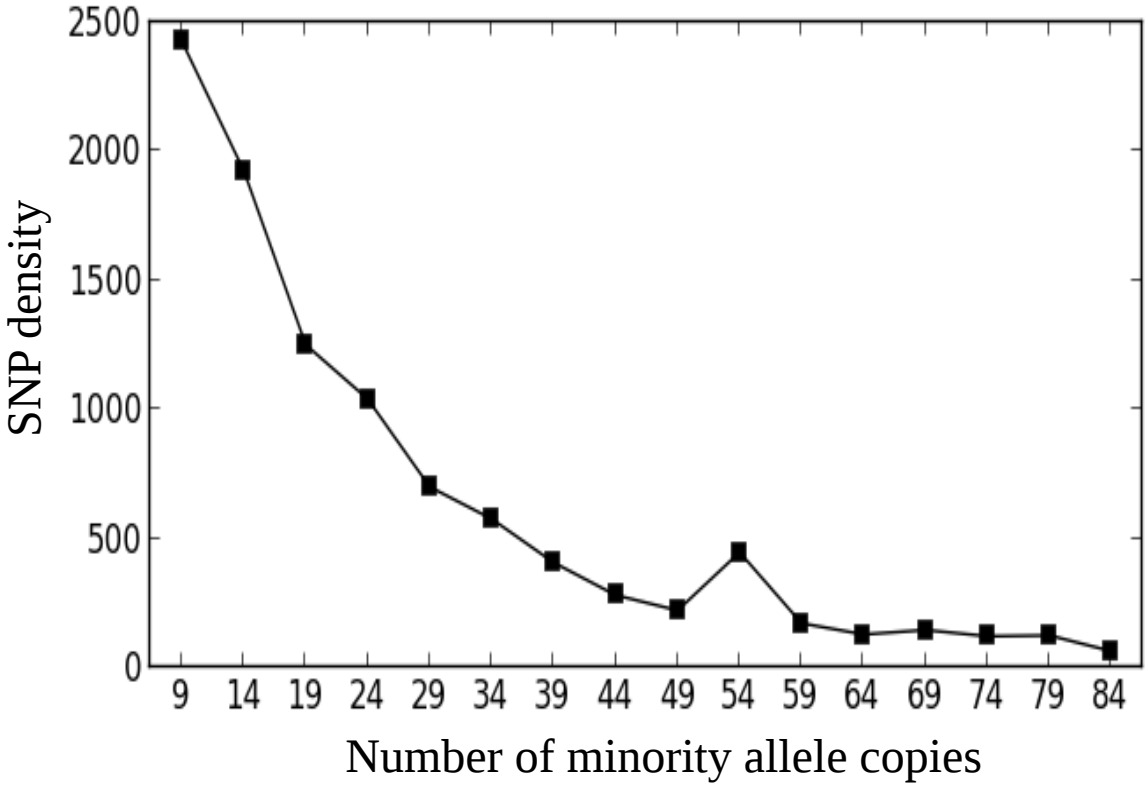


Figure 3: Site frequency spectrum representing the minority allele frequency categories based on 2-kb unlinked windows.

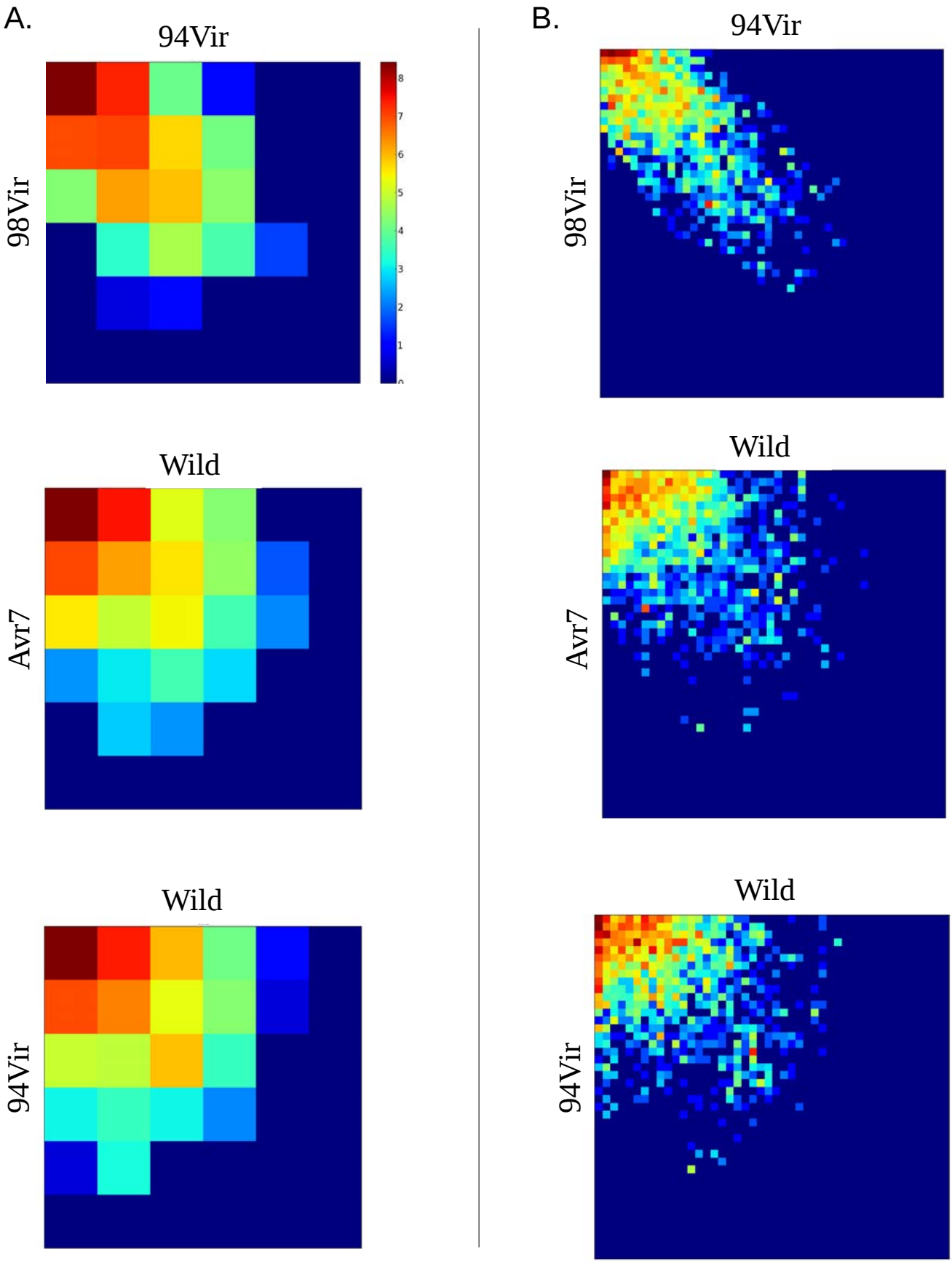


Figure 4: Joint frequency spectrum of minority allele frequency based on unlinked windows. Logarithms are used for representation. (A) Discretized (6×6 categories) JFS. (B) Raw JFS.

Parameters	Description	Statistical law	Note
N	Wild population size	normal law (standard deviation: 5000; minimum: 100; maximum: 100000)	number of diploid individuals
N1	Vir7 relative population size	normal law (standard deviation: 1; minimum: 0.01)	relative to N
N2	Avr7 relative population size	normal law (standard deviation: 2; minimum: 0.01)	relative to N
T1	time between bottleneck and first fusion	exponential law (standard deviation: 0.5; maximum: 5)	expressed in 4N generations
T2	time between first and second fusion	exponential law (standard deviation: 0.5; maximum: 10)	expressed in 4N generations
S	bottleneck strength	exponential law (standard deviation: 0.5; maximum: 5)	expressed in 4N generations
M	migration rate	normal law (standard deviation: 0.5; minimum: 0; maximum: 2)	equal to 4Nm

Table 3: Description of the parametrs used for the demographic scenarios.

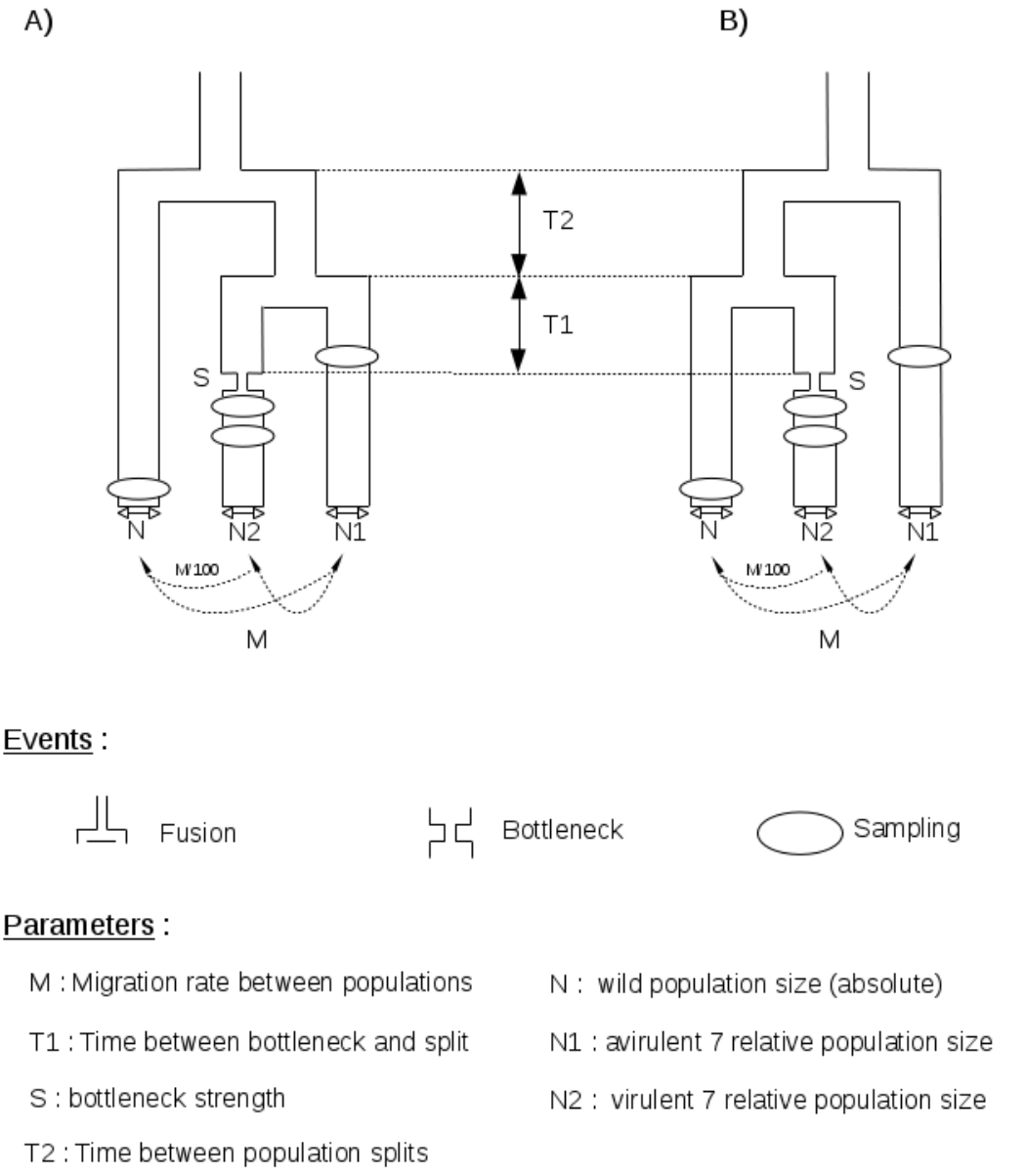


Figure 5: Two (A and B) demographic scenario used for the coalescence analysis with their events and parameters.

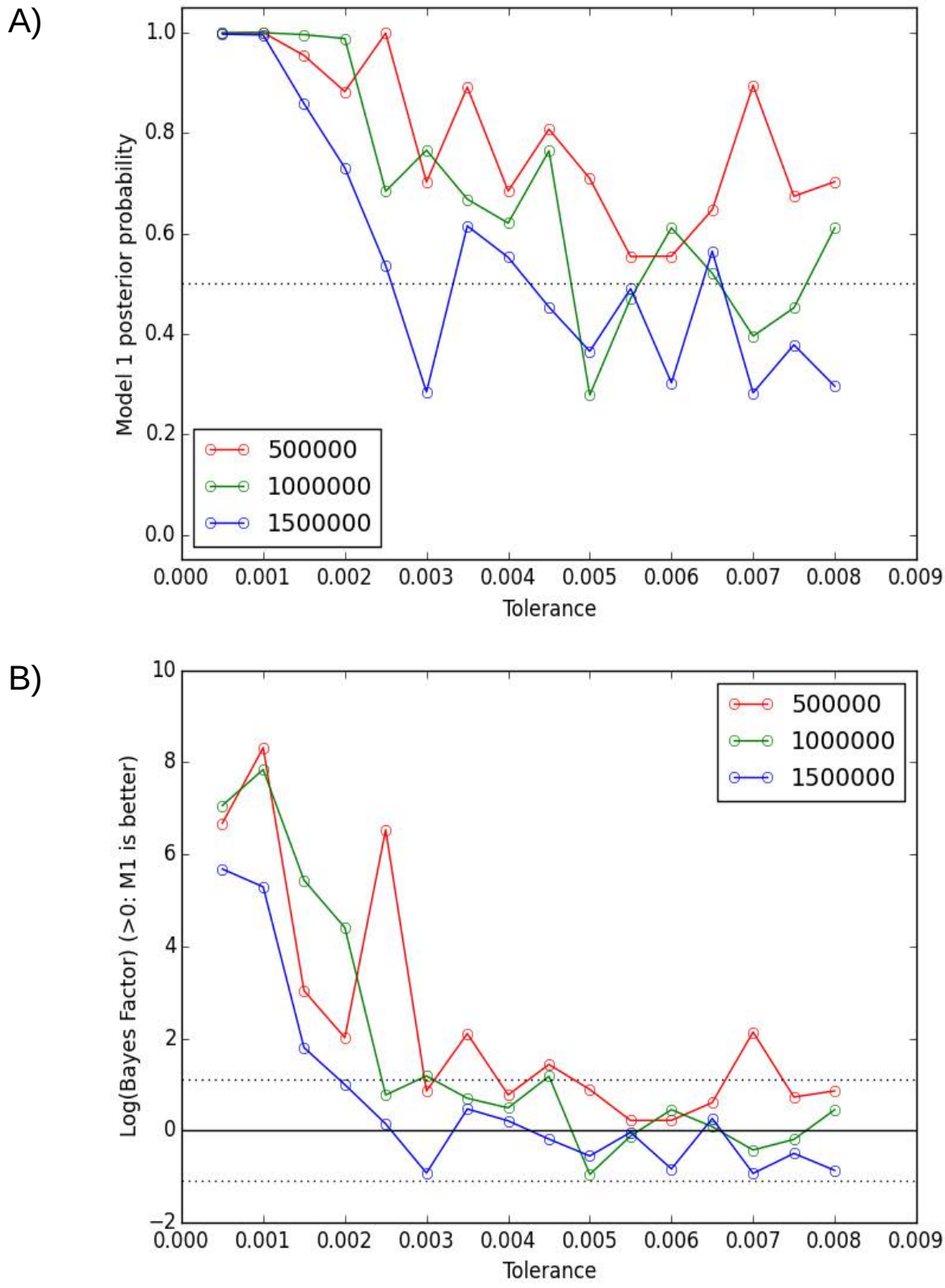


Figure 6: Model choice statistics with 250 000, 500 000 or 750 000 simulations used per model and various tolerance values. (A) Posterior probability of model A. Dotted line: 0.5 (model are equally likely). (B) Log Bayes factor of model A against B (positive values) and B against A (negative values). Dotted line: limit of Bayes factor = 3 below which each model is probably as good as the other.

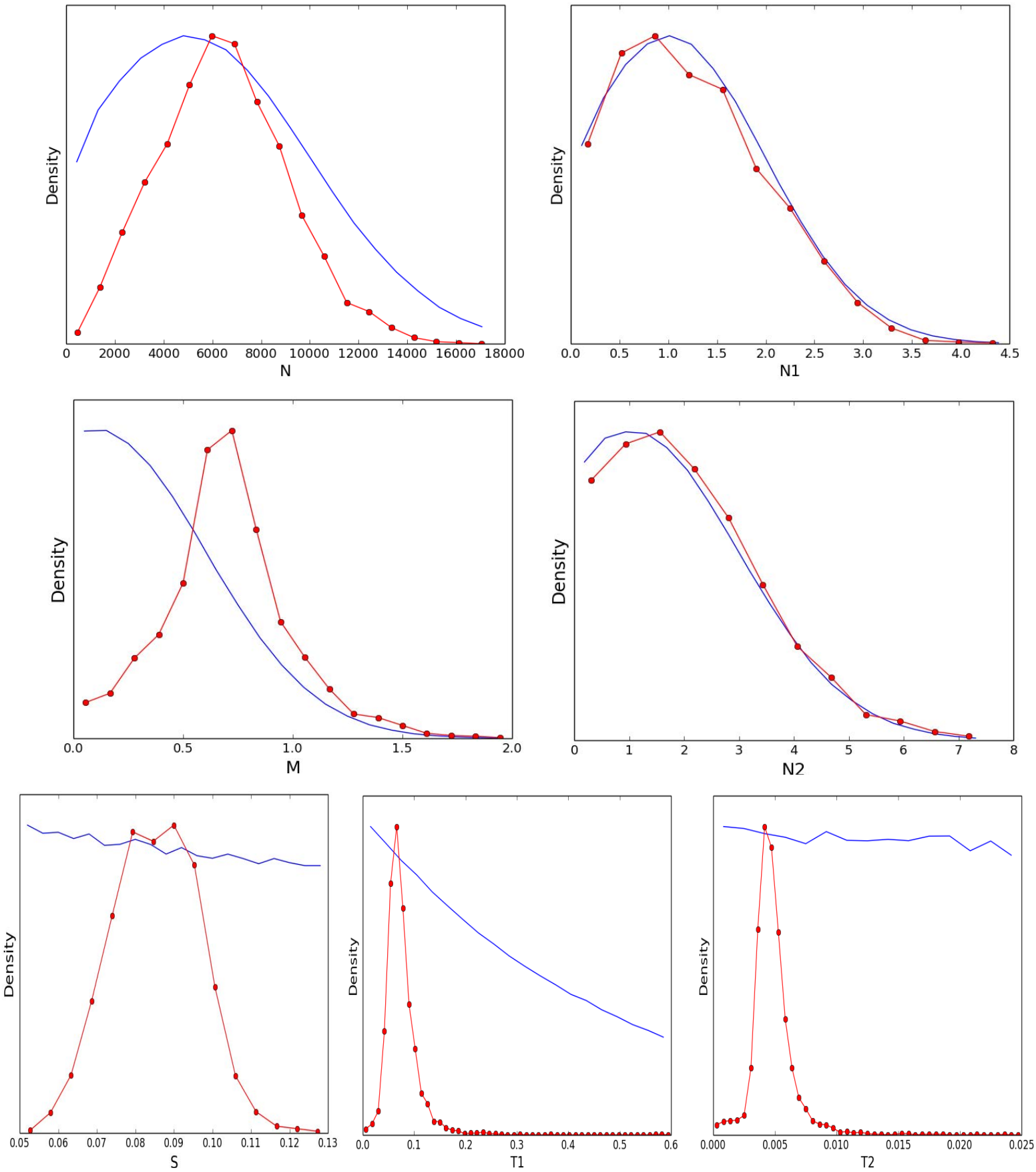


Figure 7: Demographic scenario 1 with priors (blue line) and accepted simulated points (red line and dots) distribution of the different parameters (see fig 1) for a tolerance at 0,0015. The distribution of priors S and T2 (blue line) is non apparent due to too much difference with simulated distribution.

population size are not well estimated, with a posterior distribution very similar to the prior.

We examined the pairwise correlations within the posterior distribution between pairs of parameters (data not shown). The worse (larger) coefficients of correlation were between *T1* and *M* (-0.13), *T1* and *T2* (0.08), *S* and *M* (-6%) and *T1* and *S* (5%). The correlation between *T1* (time to first split) and *M* is logical, since migration rate and time to the split have opposite effects on the time to find a common ancestor to samples taken in the two populations concerned by the split. Similarly, the bottleneck parameter can also be viewed as controlling the divergence of the virulent 7 population, which is somewhat redundant with the effect of the *S* and *T1* parameter. The correlation between the times *T1* and *T2* is less easy to interpret. However, the correlations between pairs of parameters are moderate, which is satisfactory regarding the quality of the estimation.

Genome scan

A genome scan based on *D*, *D*vir, *D*avr, F_{ST} and F_{ST} Vir-Avr was performed per linkage group for the 10 014 windows covering the whole genome. Two examples are presented in figure 8 for two regions of LG 15 and LG6 (2 Mbp each). The values of statistics for each window falling in the selected regions are showed, with the confidence intervals obtained from posterior simulation based on the adjusted ABC scenario A. The 0,1%-99.9% confidence interval is presented for Tajima's *D* and the 5%-95% confidence interval for the F_{ST}. We observed that a test based on F_{ST} alone would be more conservative than one based on Tajima's *D* alone, due to a wider distribution range for large F_{ST} values. In contrast, many observed F_{ST} are two low compared with the simulated distribution, indicating that the demographic model is still not perfectly representing neutral processes. However, we only considered F_{ST} larger than expectations, which should make our test conservative.

We automatically searched for windows with *D*vir significantly outside of expectations, and F_{ST} Vir-Avr significantly larger than expectations, as sign of a potential implication in R7 breakdown. We found 20 windows fitting our criteria (table 4). Interestingly, only two of these windows show a significant test (95%) for the global F_{ST}, indicating that, for the others, it is the virulent 7 against avirulent 7 divergence that drives genetic divergence. Two windows show a significantly negative *D*vir, indicating a putative selective sweep, the other one had a significantly positive *D*vir.

Finally, we inspected the genes contained in the 20 windows from the *M. larici-populina* genome annotation data, and found 14 genes, among which 3 are *M. larici-populina* specific (no homology found in other sequenced Pucciniales), 8 without known function and 2 coding for predicted secreted proteins (table 5).

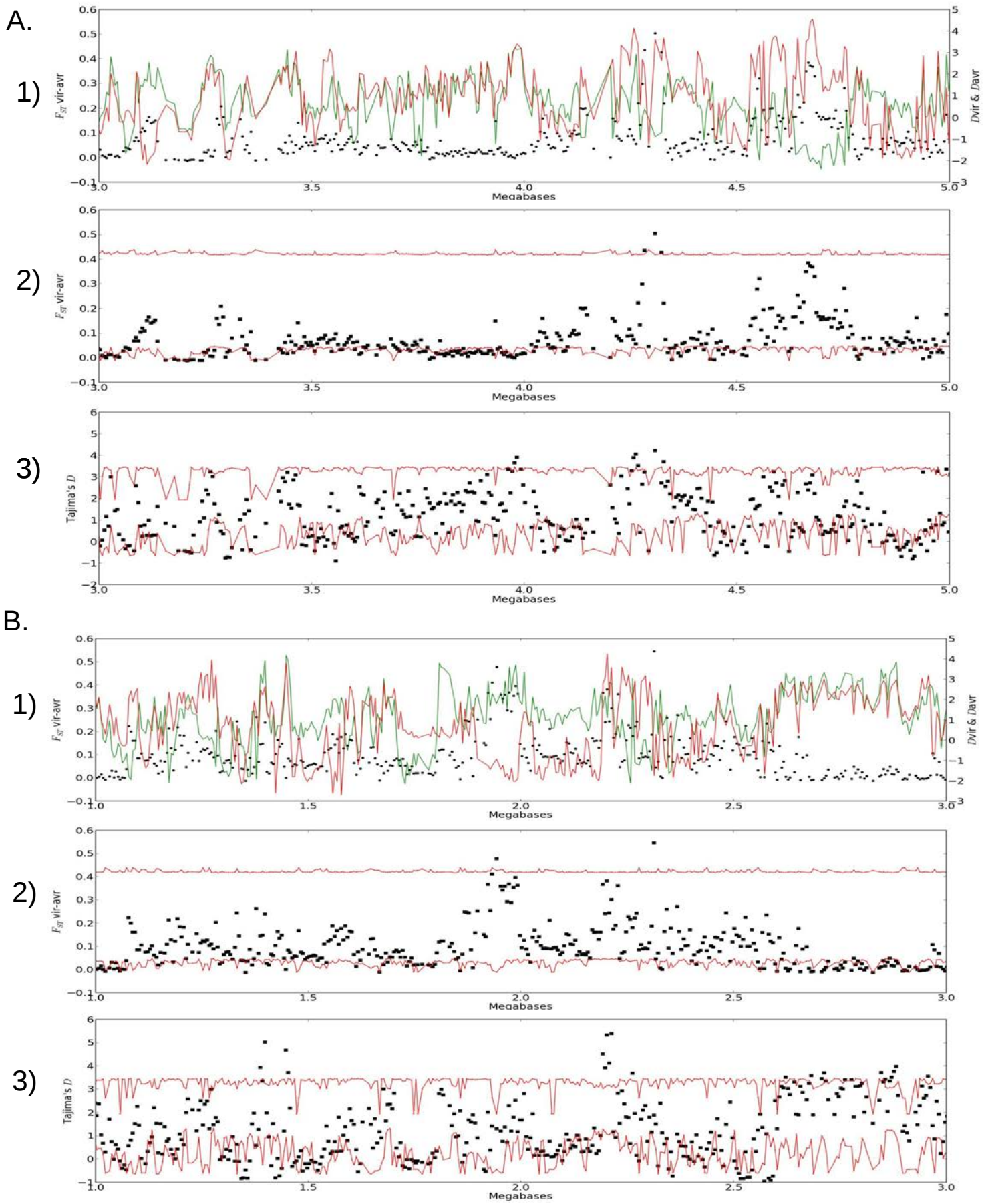


Figure 8: Genome scan of 5 kb windows for regions of LG06 (A) LG15(B) with F_{ST} Vir-Avr (black dots), *D*vir (red line) and *D*avr (green line) (1), observed F_{ST}Vir-Avr (dots) against simulations (red line) (2) and observed Tajima's *D* against simulations (red line) (3).

LG	position	SNPs	Tajima's D	significance	Tajima's Dvir	significance	Tajima's Ddavr	significance	Fst	significance	FstVir-Avr	significance
LG01	2222500,5	22	4,01	****	4,50	****	-1,13	****	0,36	ns	0,42	*
LG02	2537500,5	50	3,13	***	4,29	****	0,17	****	0,41	ns	0,48	*
LG02	2542500,5	49	3,59	****	4,27	****	-0,88	****	0,43	ns	0,52	*
LG03	4537500,5	18	3,42	****	4,39	****	-1,56	****	0,34	ns	0,44	*
LG03	4582500,5	28	3,81	****	4,20	****	-1,51	****	0,37	ns	0,45	*
LG04	5392500,5	60	4,69	****	4,41	****	-0,46	****	0,32	ns	0,43	*
LG06	4307500,5	27	4,21	****	3,61	****	-0,87	****	0,40	ns	0,50	*
LG06	4322500,5	10	2,71	*	2,63	*	-0,68	****	0,33	ns	0,43	*
LG07	4482500,5	23	3,44	****	4,57	****	-2,22	****	0,34	ns	0,42	*
LG08	647500,5	13	4,17	****	0,31	*	3,66	****	0,37	ns	0,42	*
LG08	3572500,5	22	4,21	****	3,33	****	0,25	***	0,35	ns	0,43	*
LG15	1942500,5	1	1,30	ns	-0,92	*	1,81	*	0,59	*	0,48	*
LG15	2312500,5	33	1,89	ns	-1,74	****	1,28	ns	0,48	*	0,55	*
LG16	2077500,5	40	3,77	****	3,69	****	-1,20	****	0,37	ns	0,46	*
LG16	2087500,5	18	3,86	****	3,81	****	-1,39	****	0,41	ns	0,50	*
LG16	2092500,5	10	2,90	*	3,45	****	-1,54	****	0,41	ns	0,50	*
LG16	2097500,5	7	3,49	****	3,24	****	-1,19	****	0,44	ns	0,54	*
LG16	2117500,5	2	2,22	ns	2,18	*	-0,69	*	0,42	ns	0,52	*
LG16	2292500,5	9	3,68	****	2,98	**	0,28	*	0,33	ns	0,42	*
LG16	2342500,5	3	2,78	*	2,41	*	-0,22	ns	0,39	ns	0,49	*

Table 4: Characteristics of 20 genomic regions potentially subjected to selection. LG= Linkage Group. Position represented the middle of the 5kb region. Significance value are ****>99,9%; ***>99,5%; **>99%; *>95%; ns<95%.

LG	Position	Gene_id	Function	Lenght(aa)	Homology	Secretion
LG01	2222500,5	fgenesh1_pg.C_scaffold_30000010	unknown	229	no	yes
LG04	5392500,5	estExt_fgenesh1_pg.C_220048	Rhodopsin-like receptor	1023	yes	no
LG06	4307500,5	EuGene.00210137	Unknown	266	yes	no
LG06	4307500,5	fgenesh1_pg.C_scaffold_21000136	Unknown	678	yes	no
LG06	4322500,5	estExt_fgenesh2_pg.C_210156	Unknown	562	yes	no
LG06	4322500,5	djo_fgenesh2_pg.21_162	Unknown	218	no	yes
LG07	4482500,5	EuGene.00250163	RNA binding protein	195	yes	no
LG08	647500,5	fgenesh1_pg.C_scaffold_2000062	Adenosine/AMP deaminase	360	yes	no
LG08	647500,5	EuGene.00020062	Adenosine/AMP deaminase	358	yes	no
LG08	3572500,5	fgenesh2_pg.2_503	Unknown	174	no	no
LG08	3572500,5	fgenesh1_pg.C_scaffold_2000453	Mucin Associated Surface Protein (MASP)	162	yes	no
LG15	2312500,5	fgenesh1_pg.C_scaffold_11000053	Unknown	130	yes	no
LG15	2312500,5	EuGene.00110043	Unknown	530	yes	no
LG16	2092500,5	EuGene.01100004	Rho_GTPase	269	yes	no

Table 5: Characteristics of 14 genes candidates genes. Lg= Linkage group ; Poisson represent the middle of the 5Kb region. Function, Length homology and secretion signal is obtain form the Melampsora larici-populina reference annotation at JGI.

DISCUSSION

A previous population genetic analysis has described a strong impact of the R7 breakdown on the *Melampsora larici-populina* demography (Persoons et al., in prep.). In order to understand the genetic basis of this event, we have conducted a population genomics analysis based on the resequencing of 86 isolates of *M. larici-populina* distributed 4 key samples (*Avr7*; *94Vir7*; *98Vir7*; *Wild*). Due to the strong effect of the R7 breakdown on the demography, we associated the study of the demographic consequences of this adaptive event, with the definition and comparison of two scenarios in an ABC analysis, to the study of selection with a genome scan analysis based on the best demographic scenario used as a way to generate neutral intervals. We identified 20 genomic regions of 5 kb showing signs of selective pressure that could be related with the R7 breakdown. These regions contain at least 14 genes that could be considered as candidate genes for the virulence 7.

An originality of our data set is the presence of multiple time points within our sampling (in particular, 1994 and 1998 in the virulent 7 population). Most population genetics studies are based on single time points for the obvious reason that historical samples are not available (most biological models relate to older events than in our case, and ancient DNA samples are sparse and difficult to analyze).

Tests statistics were developed to detect selection using polymorphism data based on differentiation (Beaumont and Balding, 2004; Foll and Gaggiotti, 2008; Lewontin and Krakauer, 1973; Vitalis et al., 2014), site-frequency spectrum (Kim and Stephan, 2002; Nielsen et al., 2005) and linkage disequilibrium (Ferrer-Admetlla et al., 2014; Voight et al., 2006). However, these methods can fail to detect selection under complex and/or non-equilibrium models, if the migration rate is high or if the strength selection is low (Crisci et al., 2012; Thornton et al., 2007; Vatsiou et al., 2015). The demographic and selective history of the population are likely to shape the patterns of diversity over the majority of the genome (Charlesworth, 2013). To address this problem, ABC methods can be used to both estimate the demographic history and generate a null distribution to detect genes under selection (Sunnaker et al., 2013). Due to its efficiency to handle multiple summary statistics and to accommodate more complex demographic scenarios, ABC appears as a promising framework (Bank et al., 2014). ABC is frequently used with coalescent-based simulators (due to their computational efficiency), what does not allow to directly estimate biologically realistic parameters as most parameters (such as dates and waiting times, or migration rates, or the mutation rates) are expressed as a function of the reference population size. Using a mutation rate estimate can be a way to obtain biologically meaningful values, but these estimates are not necessarily precise (De Mita et al., 2007; Veeramah et al., 2012). Using a multiple time point datasets allows to have a direct access to population size estimates thanks to the information relative

to the force of drift between successive samples in the same population (Anderson et al., 2000; Drummond and Rambaut, 2007), and to estimate jointly parameters such as selection coefficients and effective population size (Malaspinas et al., 2012).

The demographic model presenting the strongest posterior probability is model A (figure 1). In this model, the common ancestor of the virulent 7 and avirulent 7 populations is more recent (around 2100 years ago based on posterior parameter distributions) than the common ancestor of our whole sample, including the wild population (around 2300 years ago). The rate of migration estimated in our system is fairly high (a little bit less than 1 migrant per generation in total, although the total population size is not estimated with precision). The migration rate to/from the wild population was fixed to be 1% of the rate between the virulent 7 and avirulent 7 populations. This assumption was based on geographical location of populations, but can be discussed.

Under this model the population that we describe as virulent 7 had existed for a time estimated as 2100 before the bottleneck corresponding to the R7 breakdown. During this time, the population had no reason to be virulent 7 (especially since R7 did not exist for most of the time). The strength of the bottleneck is estimated to be relative moderate: the value is below 0.1, where a value of 1 would mean that no variation passed the bottleneck (note, however, that the amount of coalescence events during the bottleneck is not linearly linked to the strength of the bottleneck). This can suggest that the selective sweep was soft (the virulent 7 gene or combination of genes was already segregating in the population before the bottleneck).

However, both model selection and quantitative estimation of parameters should not be over interpreted. A scenario will always come out as the best even if none is close to reality. To increase confidence in this analysis, more scenarios (including simpler ones) should be considered to ensure that model 1 is still favored. Our model allow to provide quantitative estimation for population size and dates that are likely to be biased because of the many features that were not taken into account in the scenarios (for example, sub-structuration of populations, time-varying migration rate or population size fluctuation). Estimates should mostly be used for comparison between each other, or as a rough estimate. Although such material is not available in our case, it would be advisable to used more time distant sampling because the time gaps in our sampling are very small compared to the time scale for which the coalescent theory has been designed.

Genome scan was performed for the 18 linkage groups. Interestingly, we found many regions with high and strongly correlated *D*vir and *D*avr values and with low F_{ST} values (between 2.6 and 3 Mb for LG6, figure 8). These regions are likely to be regions of higher introgression due to gene flow between populations. F_{ST} are low, and Tajima's *D* values within populations are similar due to the homogenization of allele frequencies. Overall our genome scans revealed a fairly low F_{ST} in most of the genome with some areas presenting high F_{ST} (between 1.8 and 2 Mb for LG6, figure

5) that could correspond to barriers to introgression, representing a differentiation bloc (Bedinger et al., 2011) or could be mere effects of the neutral variance of gene genealogies across unlinked regions.

Finally we looked for genomic regions with *Dv*ir and F_{ST} vir-avr significantly out of the neutral expectation generated using the fitted demographic scenario, and we found out 20 regions with the specified characteristics (table 4). We defined our criteria based on the hypothesis that a gene involved in the R7 breakdown should drive a strong difference between virulent 7 and avirulent 7 populations, and has undergone a fast increase of frequency in the virulent 7 population, causing a significant deviation of Tajima's *D*.

Most fungal effectors protein share common features : presence of an N-terminal secretion signal, enrichment in cysteine residues, lack of functional homology in databases and small size. Such features have been widely used to determine sets of candidate effectors in the predicted proteome of fungal pathogens for which a reference genome is available (Petre et al., 2014; Lowe and Howlett, 2012). The analysis of the annotation of the candidate regions revealed 14 genes entirely or partially contained in the selected windows (table 5). Interestingly, two of them exhibited a secretion signal. They therefore likely encode secreted proteins and could be considered as candidate effectors. However, they present only part of the canonical characteristics of fungal effectors. Molecular approaches will be applied to these priority candidate effectors to validate their putative role in virulence 7 and R7 breakdown. The search for candidate genes based on the patterns of genetic variation compared to theoretical expectations might be a good complement to the search for proteins bearing a series of pre-defined characteristics that might be too restrictive to be exhaustive.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Anderson, E.C., Williamson, E.G., and Thompson, E.A. (2000). Monte Carlo evaluation of the likelihood for N(e) from temporally spaced samples. Genetics *156*, 2109–2118.

Bank, C., Ewing, G.B., Ferrer-Admettla, A., Foll, M., and Jensen, J.D. (2014). Thinking too positive? Revisiting current methods of population genetic selection inference. Trends Genet. TIG *30*, 540–546.

Beaumont, M.A., and Balding, D.J. (2004). Identifying adaptive genetic divergence among populations from genome scans. Mol. Ecol. *13*, 969–980.

Beaumont, M.A., and Nichols, R.A. (1996). Evaluating Loci for Use in the Genetic Analysis of Population Structure. Proc. R. Soc. Lond. B Biol. Sci. *263*, 1619–1626.

Bedinger, P.A., Chetelat, R.T., McClure, B., Moyle, L.C., Rose, J.K.C., Stack, S.M., van der Knaap, E., Baek, Y.S., Lopez-Casado, G., Covey, P.A., et al. (2011). Interspecific reproductive barriers in the tomato clade: opportunities to decipher mechanisms of reproductive isolation. Sex. Plant Reprod. *24*, 171–187.

Begun, D.J., Holloway, A.K., Stevens, K., Hillier, L.W., Poh, Y.-P., Hahn, M.W., Nista, P.M., Jones, C.D., Kern, A.D., Dewey, C.N., et al. (2007). Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. PLoS Biol *5*, e310.

Burke, M.K. (2012). How does adaptation sweep through the genome? Insights from long-term selection experiments. Proc Biol Sci.

Charlesworth, B. (2013). Background selection 20 years on: the Wilhelmine E. Key 2012 invitational lecture. J. Hered. *104*, 161–171.

Crisci, J.L., Poh, Y.-P., Bean, A., Simkin, A., and Jensen, J.D. (2012). Recent progress in polymorphism-based population genetic inference. J. Hered. *103*, 287–296.

Csilléry, K., François, O., and Blum, M.G.B. (2012). abc: an R package for approximate Bayesian computation (ABC). Methods Ecol. Evol. *3*, 475–479.

De Mita, S., and Siol, M. (2012). EggLib: processing, analysis and simulation tools for population genetics and genomics. BMC Genet *13*, 27.

De Mita, S., Ronfort, J., McKhann, H.I., Poncet, C., El Malki, R., and Bataillon, T. (2007). Investigation of the demographic and selective forces shaping the nucleotide diversity of genes involved in nod factor signaling in Medicago truncatula. Genetics *177*, 2123–2133.

Drummond, A.J., and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. *7*.

Duplessis, S., Bakkeren, G., and Hamelin, R. (2014). Chapter Six - Advancing Knowledge on Biology of Rust Fungi Through Genomics. In Advances in Botanical Research, Francis M. Martin, ed. (Academic Press), pp. 173–209.

Ellison, C.E., Hall, C., Kowbel, D., Welch, J., Brem, R.B., Glass, N.L., and Taylor, J.W. (2011). Population genomics and local adaptation in wild isolates of a model microbial eukaryote. Proc Natl Acad Sci U A *108*, 2831–2836.

Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V.C., and Foll, M. (2013). Robust demographic inference from genomic and SNP data. PLoS Genet. *9*.

Fay, J.C., and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. Genetics *155*, 1405–1413.

Ferrer-Admetlla, A., Liang, M., Korneliussen, T., and Nielsen, R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol. Biol. Evol. *31*, 1275–1291.

Flor, H.H. (1971). Current Status of the Gene-For-Gene Concept. Annu. Rev. Phytopathol. *9*, 275–296.

Foll, M., and Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics *180*, 977–993.

Gerard, P.R., Husson, C., Pinon, J., and Frey, P. (2006). Comparison of Genetic and Virulence Diversity of Melampsora larici-populina Populations on Wild and Cultivated Poplar and Influence of the Alternate Host. Phytopathology *96*, 1027–1036.

Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., and Bustamante, C.D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. *5*.

Haudry, A., Cenci, A., Ravel, C., Bataillon, T., Brunel, D., Poncet, C., Hochu, I., Poirier, S., Santoni, S., Glemin, S., et al. (2007). Grinding up wheat: a massive loss of nucleotide diversity since domestication. Mol. Biol. Evol. *24*, 1506–1517.

Hudson, R.R., Slatkin, M., and Maddison, W.P. (1992). Estimation of levels of gene flow from DNA sequence data. Genetics *132*, 583–589.

Jones, J.D., and Dangl, J.L. (2006). The plant immune system. Nature *444*, 323–329.

Kim, Y., and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics *160*, 765–777.

Lewontin, R.C., and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the

selective neutrality of polymorphisms. Genetics *74*, 175–195.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. *25*, 2078–2079.

Li, M., Tian, S., Yeung, C.K.L., Meng, X., Tang, Q., Niu, L., Wang, X., Jin, L., Ma, J., Long, K., et al. (2014). Whole-genome sequencing of Berkshire (European native pig) provides insights into its origin and domestication. Sci. Rep. *4*.

Liti, G., Carter, D.M., Moses, A.M., Warringer, J., Parts, L., James, S.A., Davey, R.P., Roberts, I.N., Burt, A., Koufopanou, V., et al. (2009). Population genomics of domestic and wild yeasts. Nature *458*, 337–341.

Lowe, R.G.T., and Howlett, B.J. (2012). Indifferent, affectionate, or deceitful: lifestyles and secretomes of fungi. PLoS Pathog. *8*.

Malaspinas, A.-S., Malaspinas, O., Evans, S.N., and Slatkin, M. (2012). Estimating allele age and selection coefficient from time-serial data. Genetics *192*.

Mardis, E.R. (2008). Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet *9*, 387–402.

McDowell, J.M. (2011). Genomes of obligate plant pathogens reveal adaptations for obligate parasitism. Proc. Natl. Acad. Sci. U. S. A. *108*, 8921–8922.

Namroud, M.C., Beaulieu, J., Juge, N., Laroche, J., and Bousquet, J. (2008). Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. Mol Ecol *17*, 3599–3613.

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G., and Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. Genome Res *15*, 1566–1575.

Pernaci, M., De Mita, S., Andrieux, A., Petrowski, J., Halkett, F., Duplessis, S., and Frey, P. (2014). Genome-wide patterns of segregation and linkage disequilibrium: the construction of a linkage genetic map of the poplar rust fungus Melampsora larici-populina. Front. Plant Sci. *5*.

Persoons, A., Morin, E., Delaruelle, C., Payen, T., Halkett, F., Frey, P., De Mita, S., and Duplessis, S. (2014). Patterns of genomic variation in the poplar rust fungus Melampsora larici-populina identify pathogenesis-related factors. Front. Plant Sci. *5*.

Persoons, A., Fabre, B., Frey, P.,De Mita, S.and Halkett, F (in preparation). Population replacment following a major selection event in the plant pathogen *Melampsora larici-populina*.

Petre B, Joly DL, Duplessis S (2014) Effector proteins of rust fungi. *Front Plant Sci.* 5:416

Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W., et al. (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Res. *19*, 826–837.

Pinon, J., and Frey, P. (1997). Structure of Melampsora larici-populina populations on wild and cultivated poplar. Eur. J. Plant Pathol. *103*, 159–173.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature *419*, 832–837.

Steenackers, J., Steenackers, M., STEENACKERS, V., and Stevens, M. (1996). Poplar diseases, consequences on growth and wood quality. Biomass Bioenergy *10*, 267–274.

Sunnaker, M., Busetto, A.G., Numminen, E., Corander, J., Foll, M., and Dessimoz, C. (2013). Approximate Bayesian computation. PLoS Comput. Biol. *9*.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585–595.

Thornton, K.R., Jensen, J.D., Becquet, C., and Andolfatto, P. (2007). Progress and prospects in mapping recent selection in the genome. Heredity *98*, 340–348.

Vatsiou, A.I., Bazin, E., and Gaggiotti, O.E. (2015). Detection of selective sweeps in structured populations: a comparison of recent methods. Mol. Ecol.

Veeramah, K.R., Wegmann, D., Woerner, A., Mendez, F.L., Watkins, J.C., Destro-Bisol, G., Soodyall, H., Louie, L., and Hammer, M.F. (2012). An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. Mol. Biol. Evol. *29*, 617–630.

Vitalis, R., Gautier, M., Dawson, K.J., and Beaumont, M.A. (2014). Detecting and measuring selection from gene frequency data. Genetics *196*.

Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. PLoS Biol. *4*.

Wang, M., Yu, Y., Haberer, G., Marri, P.R., Fan, C., Goicoechea, J.L., Zuccolo, A., Song, X., Kudrna, D., Ammiraju, J.S.S., et al. (2014). The genome sequence of African rice (Oryza glaberrima) and evidence for independent domestication. Nat. Genet. *46*, 982–988.

Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. Theor Popul Biol *7*, 256–276.

Xhaard, C., Andrieux, A., Halkett, F., and Frey, P. (2009). Characterization of 41 microsatellite loci developed from the genome sequence of the poplar rust fungus, Melampsora larici-populina. Conserv. Genet. Resour. *1*, 21–25.

Xhaard, C., Fabre, B., Andrieux, A., Gladieux, P., Barres, B., Frey, P., and Halkett, F. (2011). The genetic structure of the plant pathogenic fungus Melampsora larici-populina on its wild host is extensively impacted by host domestication. Mol Ecol *20*, 2739–2755.

Zeng, K., Shi, S., and Wu, C.-I. (2007). Compound tests for the detection of hitchhiking under positive selection. Mol. Biol. Evol. *24*, 1898–1908.

Chapitre 5

Discussion générale

Chapitre 5 - Synthèse et conclusion générale

1. Le contournement de la résistance 7, un bouleversement populationnel

1.1. Structuration démographique

Les objectifs principaux de ce travail de thèse étaient de décrire l'effet démographique du contournement de la résistance 7 et d'en identifier les déterminants génétiques. L'impact démographique de ce contournement de résistance a été très fort et il s'est accompagné d'un bottleneck et d'une expansion démographique du groupe génétique constitué par les individus porteurs de la virulence 7 (chapitre 3 et 3bis). Étant donné cet effet fort et le fait que la sélection et la démographie façonnent les génomes en synergie (Charlesworth, 2013), nous les avons étudiées conjointement afin de détecter des régions génomiques potentiellement impliquées dans ce contournement de résistance (chapitre 4).

L'étude de génétique des populations du chapitre 3 a permis de mettre à jour trois groupes génétiques au sein des isolats échantillonnés et d'en décrire les caractéristiques. Deux de ces groupes avaient déjà été identifiés lors d'une précédente étude (Xhaard et al., 2011).

Le groupe des avirulents 7 a disparu à partir de 1997 en France, vraisemblablement remplacé par les virulents 7 (Persoons et al., in prep). Cependant notre étude ne porte que sur des isolats français, et il pourrait être intéressant d'étudier les groupes génétiques à une échelle plus large pour déterminer si ce groupe génétique n'a pas subsisté ailleurs en Europe. On peut imaginer qu'il existe des régions relativement isolées où la pression de sélection causée par le déploiement de la résistance 7 était moins forte voire inexistante. Une étude complémentaire a été menée lors du stage de Master 1 de Marius Colin (stage que j'ai co-encadré) sur 142 isolats européens et canadiens échantillonnés en 2003 et génotypés à l'aide des mêmes marqueurs microsatellites que l'étude Barres et al., (2008) Ces isolats ont été ajoutés aux 600 utilisés lors de l'analyse rapportée dans le chapitre trois et une étude d'assignation bayésienne a été réalisée. Le nombre de clusters expliquant au mieux les données est de cinq. Les deux groupes supplémentaires correspondent à des isolats canadiens et à ceux décrits comme asexués par Xhaard et al., (2011) lesquels forment chacun un groupe génétique distinct. Les trois groupes génétiques décrits dans le chapitre 3 sont retrouvés au sein des isolats européens de 2003. Cette étude a montré que les individus sauvages se retrouvent dans le Sud de la France et de l'Europe principalement, les individus virulents 7 dans l'Ouest de l'Europe et les individus avirulents 7 dans l'Est (figure 1). Ainsi, ce dernier groupe génétique aurait disparu de France où il aurait été remplacé par le groupe génétique des virulents 7 (du fait de la forte pression de sélection due à la plantation massive des

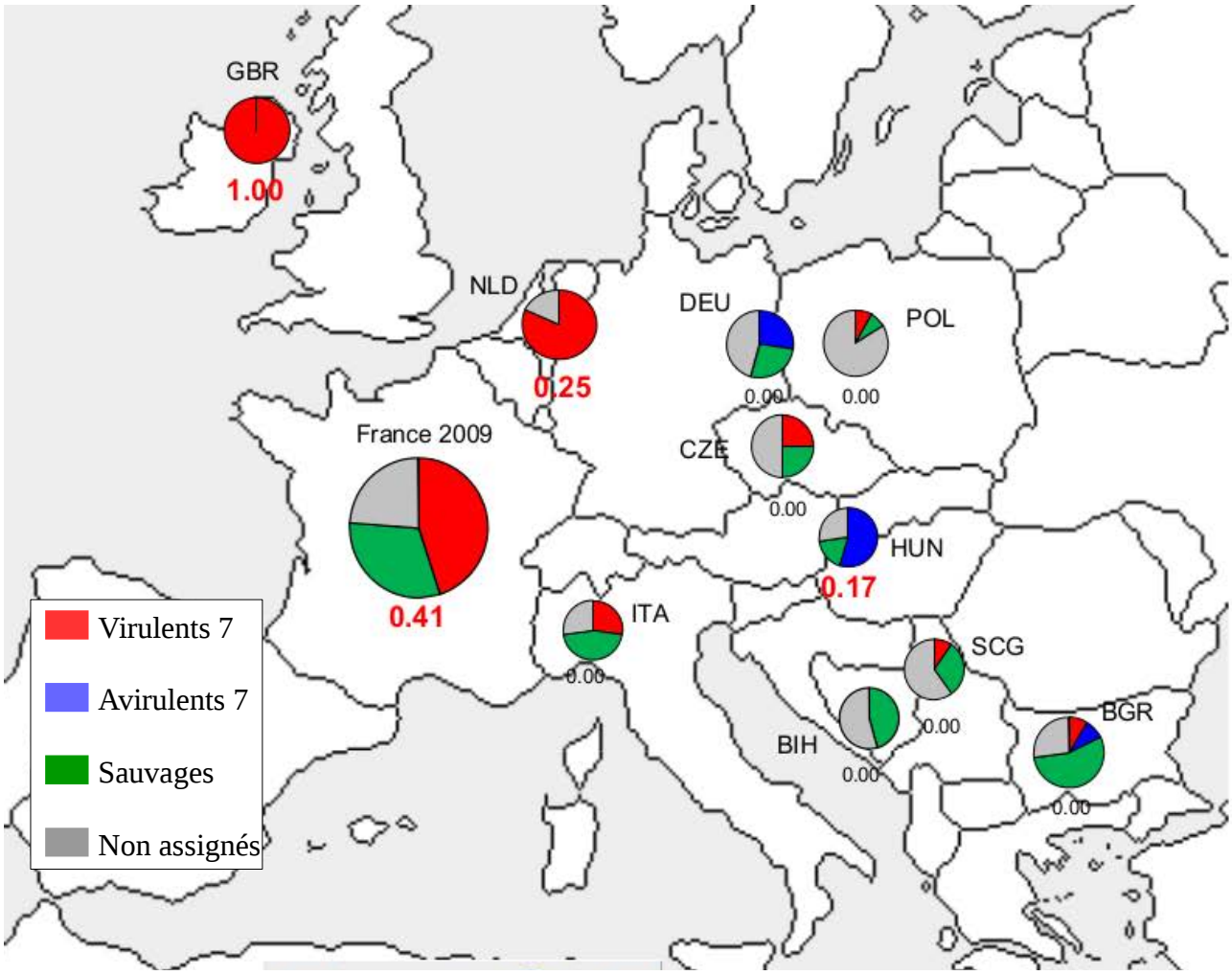


Figure 1 : Distribution géographique des groupes génétiques, à l'échelle européenne, montrant la proportion d'individus assignés (coefficient d'assignation supérieur à 0,8) aux groupes des virulents 7 (en rouge), avirulents 7 (en bleu) et sauvages (en vert). La proportion d'individus non assignés (aucune probabilité supérieure à 0,8) est indiquée en gris. Le taux d'individus présentant la virulence 7 (testé par pathotypage) est indiqué en dessous des graphiques pour chaque pays où l'information est disponible. Le diamètre des disques est proportionnel au nombre d'échantillons utilisés.

peupliers porteurs de la résistance 7) mais il aurait subsisté dans d'autres régions d'Europe. Ces conclusions sont toutefois à nuancer de part la forte proportion d'individus non assignés (figure 1) qui indiquent qu'une partie de la structuration populationnelle n'est probablement pas bien représentée par cette analyse.

Ces conclusions sur l'histoire démographique de *M. larici-populina* associée au contournement de la résistance 7 sont conditionnées par le modèle démographique inféré dans le chapitre 4. Or cette inférence est forcément dépendante des modèles testés. Les analyses par ABC peuvent assigner une très forte probabilité postérieure à un modèle même s'il est très éloigné de la réalité (il suffit que tous autres modèles testés le soient également). Il est difficile d'exclure la possibilité qu'un modèle non pris en compte dans l'analyse aurait pu mieux expliquer nos données. Par exemple, l'hypothèse d'une origine des virulents 7 par une hybridation entre les avirulents 7 et les sauvages aurait pu paraître séduisante d'un point de vue biologique. Cependant les résultats des assignations Bayésiennes du chapitre 2 indiquent très clairement une différenciation forte de ces deux groupes génétiques. De plus, le modèle démographique du chapitre 4 indique un taux de migration assez fort entre ces deux groupes. Ainsi tout indique que la virulence 7 serait apparue dans le groupe des virulents 7 (qui n'a forcément porté la virulence 7 depuis son origine et a pu l'acquérir par mutation) qui aurait alors subi un goulot d'étranglement et une expansion démographique suivi de flux de gènes très importants avec le groupe des avirulents 7 initialement en place.

Le fait que le groupe virulent 7 soit retrouvé en 2003 dans d'autres pays européens est cohérent avec le temps de fusion long entre les groupes virulent 7 et avirulent 7 (estimé à 2100 ans selon l'analyse ABC du chapitre 4) puisque ce groupe pourrait avoir une origine ailleurs en Europe et donc n'être apparu en France en 1994 que grâce à un évènement de dispersion. De ce fait, la question de l'apparition de la virulence 7 reste ouverte. Cette virulence pourrait avoir pré-existé dans le groupe ancêtre des virulents 7 sans que la résistance correspondante n'existe chez le peuplier (comme c'est le cas de la virulence dite X décrite dans la partie 5.3). Sous cette hypothèse, l'apparition des virulents 7 en France, en 1994, serait donc due à un évènement de migration plutôt que de mutation. Ce scénario pourrait expliquer l'intensité modérée du goulot d'étranglement chez les virulents 7 (démontrée par le niveau conséquent de diversité détecté dans les premières années dans cette population), car la virulence 7 pourrait avoir ségrégé suffisamment longtemps dans la population ancestrale pour pouvoir être associée à un fond génétique diversifié (*soft selective sweep* ; Hermisson and Pennings, [2005]).

Désormais les clones de peupliers porteurs de la résistance 7 sont beaucoup moins plantés, passant de 1 200 000 en 1996 à moins de 200 000 en 2006 (Hayden et al., non publié), ce qui réduit la pression de sélection sur la virulence 7. De manière cohérente, des individus assignés au groupe

virulent 7 mais ne portant pas la virulence 7 (testé par pathotypage) sont apparus ces dernières années. Une approche de génomique comparative entre ces isolats et des individus phénotypiquement virulents 7 pourrait être intéressante car, les individus étant issus du même fond génétique, le niveau de différenciation génétique global serait grandement réduit. De ce fait la recherche des bases moléculaires de la virulence 7 serait plus simple, alors que l'histoire démographique de cette population rend la détection du gène causal plus difficile.

1.2. Scénarios démographiques

L'étude de recherche des traces de la sélection du chapitre 4 est basée sur un modèle démographique qui est primordial puisque ses valeurs de paramètres conditionnent les distributions attendues des statistiques utilisées dans les tests. Bien que le scénario démographique du chapitre 4 ait permis de préciser les liens historiques entre ces groupes, l'estimation des paramètres et en particulier des dates pose question. Du fait des simplifications des modèles utilisés, l'estimation quantitative des paramètres biologiques peut présenter des biais (ce qui est problématique notamment pour les interprétations concrètes). Un des intérêts de notre échantillonnage est la présence de plusieurs points historiques. Dans la majorité des analyses par ABC, il n'y a qu'un seul point temporel et ainsi l'estimation des temps des modèles est effectuée à partir du taux de mutations (μ) : La taille efficace *N*e (4*N*eμ) et les temps T (exprimés en 4*Ne* générations) sont déduits d'une estimation de θ grâce à une estimation parfois grossière de μ par génération (Veeramah et al., 2012). L'utilisation de plusieurs points historiques permet d'avoir accès à une estimation directe de la taille efficace et donc de temps (Anderson et al., 2000; Drummond and Rambaut, 2007). Étant donné la relative faiblesse du signal de l'estimation de *N*e (variance postérieure très importante) du chapitre 4, notre estimation des temps est sans doute assez peu fiable puisque dépendante de *N*e, et cela même si l'estimation des paramètres *T1* et *T2* sont relativement mieux estimés. Cela est sans doute dû à un manque de signal du fait d'un écart assez faible entre les dates d'échantillonnages utilisées (une dizaine d'années). De plus, nos modèles sont par nature trop simplistes et peuvent donc biaiser l'estimation de *N*e (Bank et al., 2014).

Plusieurs solutions à court, moyen et long terme sont envisagées pour augmenter la qualité des inférences du scénario démographique :

- A court terme, 20 millions de simulations des paramètres sous chacun des deux modèles (au lieu des 1,5 millions du chapitre 4) seront utilisées pour le choix de modèle et pour en ajuster les paramètres plus finement. L'objectif est principalement de voir s'il est possible d'obtenir plus de signal pour les tailles de populations *N1* et *N2* principalement. Pour le choix du modèle, les statistiques montrent que l'augmentation du nombre de données a tendance à baisser les probabilités postérieures (chapitre 4, figure 3), il est donc peu probable que nous augmentions la significativité

avec 20 millions de simulations.

- A moyen terme, d'autres scénarios seront ajoutés aux deux déjà testés, dont un scénario avec des hybridations entre populations dans le but de prendre en compte une apparition de la virulence 7 par hybridation. Divers scénarios seront considérés en faisant varier les évènements démographiques et les paramètres (pas de goulot d'étranglement, migration entre avirulents 7 et sauvages égale aux autres taux, population non échantillonnée réalisant des flux de gènes avec les différentes populations) dans le but de vérifier la convergence vers le scénario 1.

- A long terme, l'utilisation de données temporelles bien plus larges telles que celles utilisées dans les analyses d'ADN ancien (Devault et al., 2014) pourrait se révéler très intéressante dans l'inférence de l'histoire démographique de *M. larici-populina*. En effet l'écart de temps entre les échantillons rapportés aux temps de fusion des populations est finalement assez faible (10 ans contre 2000 ans) pouvant également biaiser l'estimation des tailles de population et donc des temps de fusion de populations.

2. Vers l'identification des déterminants génétiques

2.1. Identification de gènes candidats par la génomique comparative

Le chapitre 2 s'est concentré sur la recherche de gènes candidats d'avirulence, par une approche d'identification de polymorphismes non-synonymes au sein de gènes codant des protéines sécrétées, considérées comme effecteurs candidats. Plusieurs types de polymorphismes ont été caractérisés dans des gènes codant des effecteurs pour leurs effets sur la virulence d'agents pathogènes. L'effecteur apoplastique *Avr*4 de *Cladosporium fulvum*, pathogène de la tomate, échappe à la reconnaissance par la protéine de résistance Cf4 par des changements d'acides aminés (Joosten et al., 1994). L'effecteur AvrL567 de *M. lini* échappe à la reconnaissance des récepteurs L5, L6 et L7 de *Linum usitatissimum* par une série de mutations non-synonymes (Dodds et al., 2006). De plus, la famille d'effecteurs Avr3a présente dans plusieurs espèces de *Phytophthora* (dont *P. infestans* et *P. sojae*), présente des mutations non-synonymes permettant à ces oomycètes d'échapper à la reconnaissance par les différentes plantes hôtes (Boutemy et al., 2011; Yaeno et al., 2011). Cependant, au-delà des mutations ponctuelles sur lesquelles j'ai concentré mon travail, plusieurs autres mécanismes moléculaires peuvent expliquer l'émergence d'une virulence.

L'apparition de nouveaux gènes et de nouvelles fonctions par insertion est également un mécanisme connu de virulence. Le transfert horizontal permet l'acquisition de nouveaux gènes issus d'espèces phylogénétiquement proches ou non. Ce mécanisme a d'abord été décrit chez les bactéries, principalement pour le transfert de gènes de résistance aux antibiotiques (Akiba et al., 1960). Plus récemment, des transferts horizontaux ont été décrits chez des parasites des plantes et

cela représenterait une force évolutive non négligeable de ces espèces (Danchin et al., 2010; Soanes and Richards, 2014). Différentes études ont permis de mettre à jour 46 transferts de gènes vers des micro-organismes pathogènes des plantes, avec des gènes codant des protéines impliquées dans la pathogénie soit codant pour des fonctions pathogènes putatives (Soanes and Richards, 2014). Par exemple, le gène de la toxine *ToxA* a subi un transfert horizontal récent, d'un champignon pathogène du blé, *Stagonospora nodorum*, à un autre, *Pyrenophora tritici-repentis* (Friesen et al., 2006). Cet évènement unique est censé s'être produit juste avant 1941 et être responsable de l'émergence de l'helminthosporiose du blé dans les années 1940.

Les délétions sont également connues pour promouvoir la virulence (Raffaele and Kamoun, 2012). Des délétions dans les régions génomiques portant les gènes d'avirulence *AvrLm*1 et *AvrLm*6 chez *Leptosphaeria maculans* ont rendu les lignées correspondantes virulentes sur les plants de *Brassica napus* porteurs des gènes de résistance *RLM1* et *RLM6* (Fudal et al., 2009; Gout et al., 2007). La grande concentration de rétrotransposons dans les régions riches en isochores AT, lesquelles abritent des gènes codant des effecteurs chez *L. maculans*, peut avoir contribué à l'augmentation de la fréquence des évènements de recombinaison, conduisant à des délétions. Le gène *AvrLm6* se situant en amont d'éléments transposables dégénérés et des mutations RIP engendrant des codons stop prématurés ayant été détectés dans les allèles d'*AvrLm6*, il est envisagé que la virulence engendrée par ce gène soit due à des délétions inactivant la reconnaissance du produit du gène (Fudal et al., 2009; Van de Wouw et al., 2010). Différentes études ont démontré par génétique les interactions gène-pour-gène dans le pathosystème *L. maculans-B. napus* où les gènes d'avirulence fongique (*AvrLm*) sont la cible des gènes de résistance de la plante (*Rlm*) (Rouxel and Balesdent, 2005). Si l'on applique les principes de la relation gène-pour-gène, on peut conclure que l'absence des gènes d'avirulence, qui sont les cibles des gènes de résistance correspondants chez la plante, empêche la mise en place de l'immunité spécifique de la plante et permet le succès de l'infection. D'autres exemples de délétions favorisant la virulence sont connus, tels que la délétion du gène d'avirulence *Avr-Pita* chez *M. oryzae* (Dai et al., 2010; Orbach et al., 2000) ou celle de l'effecteur de type RXLR Avr4 de *Phytophthora infestans* (van Poppel et al., 2008). Dans le chapitre 2, la recherche des régions délétées (détectées parce qu'elles ne présentent aucune couverture) a permis de mettre à jour des régions potentiellement impliquées dans le système d'appariement (Persoons et al., 2014). Pour aller plus loin dans la détection des délétions il faudrait réaliser un assemblage du génome de chacun des individus échantillonnés, permettant ainsi la détection des régions absentes chez les individus virulents. En effet, notre approche est basée sur l'alignement de toutes les séquences générées sur un génome de référence provenant d'un individu virulent 7, ce qui fait que seules les régions présentes chez cet individu sont considérées. Cependant, séquencer à une profondeur permettant un assemblage de qualité suffisante représenterait un coût conséquent.

L’acquisition d'un nouveau génome de référence avirulent 7 permettrait de détecter les insertions et délétions potentiellement associées à l'émergence de la virulence. Au cours de ma thèse, j'ai séquencé le génome de l'isolat 93GS3 qui est avirulent 7. Ce séquençage *de novo* a été réalisé avec deux banques différentes de séquençage Illumina, une librairie paired-end générant des séquences courtes à partir de fragments courts d'ADN (500 bases environ) et une librairie mate pair générant séquences de taille similaire à partir de fragments d'ADN plus longs (> 5000 kilobases). Un premier essai d'assemblage *de novo* a été réalisé à l'aide du programme SOAP *de novo* (Luo et al., 2012b), en utilisant la librairie paired-end pour l'assemblage et la librairie mate pair pour le scaffolding (utilisation des longues lectures pour joindre les régions assemblées entre elles). L'obtention d'un premier assemblage de ce nouveau génome de référence de plus de 11 000 scaffolds nécessite d'être améliorée en optimisant les paramètres d'assemblage *de novo*. De plus, le transcriptome de cet isolat a été séquencé dans deux conditions distinctes (spores et lésions sporulantes 4 jours après inoculation) par RNA sequencing. Cela permettra de valider l'expression des gènes une fois l’assemblage finalisé, et de différencier les gènes exprimés de façon identique dans les deux conditions testées de ceux exprimés uniquement dans les feuilles infectées et donc potentiellement impliqués le processus infectieux. Ce nouveau génome de référence avirulent 7 permettra aussi d'aligner les séquences Illumina des isolats avirulents 7 décrites dans le chapitre 2 sur celui-ci et ainsi de détecter d'éventuels réarrangements chromosomiques, de larges insertions ou délétions, qui pourraient être associés à la virulence 7 si cette dernière ne s'expliquait pas par des mutations de type SNP.

2.2. Identification de gènes candidats par la génomique des populations

Le chapitre 4 traite de la recherche de gènes candidats impliqués dans le contournement de la résistance 7, par une approche de scan génomique. Cette approche, contrairement à la précédente (chapitre 2), peut être considérée comme naïve puisqu’elle ne s'intéresse pas aux fonctions des gènes avant de les avoir identifiés. Ce type d'analyse est de plus en plus utilisé avec l'accès facilité aux séquences (Shendure and Ji, 2008). Les méthodes de scan génomique permettent de s’intéresser à de très nombreux phénomènes biologiques tels que la divergence entre populations (Beaumont and Balding, 2004; Begun et al., 2007; Liti et al., 2009), leur histoire comme par exemple la domestication (porc : Li et al., 2014, blé :Haudry et al., 2007, riz : Wang et al., 2014), l'adaptation (Burke, 2012; Ellison et al., 2011; Namroud et al., 2008), l'invasion (Lombaert et al., 2014; Stukenbrock et al., 2011), et dans notre cas, les balayages sélectifs (De Mita et al., 2007; Nielsen et al., 2005; Pariset et al., 2009). Concernant les balayages sélectifs, il s’agit d'identifier les gènes sous pression de sélection positive au sein des génomes, idéalement après avoir décrit l'histoire démographique des isolats pour en éviter les effets confondants (Cadzow et al., 2014). Ce type

d'analyse a été réalisé avec succès sur de nombreuses espèces modèles. Par exemple, une étude de génomique des populations basée sur des scans génomiques étudiant le déséquilibre de liaison, la diversité nucléotidique et les fréquences des haplotypes entre populations humaines a permis de mettre à jour des gènes sous sélection dans les différentes populations, tels que des gènes de résistance à la malaria (Kimura et al., 2007). Une analyse approfondie de la structure populationnelle de *Saccharomyces cerevisiae* et *S. paradoxus* à l'aide du séquençage d'une trentaine individus par espèce (Liti et al., 2009) a permis l'identification et l'étude de gènes de réparation de l'ADN impliqués dans la variation du taux de mutation des espèces, générant des souches avec un taux de mutation très élevés (Demogines et al., 2008).

Dans notre cas, l'étude et la description des populations de *M. larici-populina* présentées au chapitre 3 a permis d'identifier un scénario démographique et de l'utiliser pour trouver des gènes sous sélection à l'aide d'un scan génomique. Mais ce type de méthode peut manquer de puissance pour détecter les balayages sélectifs trop récents (où l'allèle favorable n'est pas allé jusqu'à la fixation complète). La puissance (le taux de vrais positifs détectés) du *D* de Tajima, par exemple, n'excède pas 0,5 et est même beaucoup plus faible si le goulot d'étranglement est faible ou trop récent (Depaulis et al., 2003). C'est dans ce cadre que des méthodes plus sophistiquées sont apparues, en particulier pour l'humain, avec pour objectif de détecter plus finement les balayages sélectifs récents. Il existe plusieurs catégories de méthodes statistiques pour détecter les balayages sélectifs, qui peuvent être basées sur différents types de signaux (Akey, 2009). Certaines sont basées sur le spectre de fréquences alléliques, soit sur la base de statistiques comme le *D* de Tajima et ses variants qui en font un résumé, ou sur la base d'une comparaison de la vraisemblance des données (exploitant donc le jeu de données de manière plus complète) sous des modèles avec et sans sélection (Kim and Stephan, 2002). D'autres méthodes exploitent l'extension du déséquilibre de liaison en cas d'augmentation en fréquence récente d'un allèle, telles que l'homozygotie étendue des haplotypes (EHH) qui est basée la longueur des haplotypes en forte fréquence (Sabeti et al., 2002). Enfin, des méthodes utilisent le niveau de différenciation génétique entre populations pour déterminer des régions qui présentent un excès, ou un déficit de différenciation (Bazin et al., 2010; Beaumont and Balding, 2004). Ces régions peuvent être le signe d'adaptation locale à des conditions variables entre populations, ou d'un balayage n'ayant eu lieu que dans une seule des populations. Certaines méthodes combinent plusieurs sources d'information, telle une variante de l'EHH prenant en compte le contraste entre différentes populations (Tang et al., 2007). Dans le chapitre 4 nous avons essayé de combiner le spectre de fréquence allélique et la différenciation entre populations en utilisant des statistiques assez simples (*D* de Tajima et F_{ST}). Exploiter l'information du déséquilibre permettrait d'augmenter la puissance de détection, notamment pour des balayages sélectifs récents, ce qui est le cas du contournement de la résistance 7. Néanmoins,

cela nécessiterait l'identification des haplotypes du génome de référence issu d'un individu dicaryotique, autrement dit la reconstruction de l'association des états alléliques le long des brins haplotypiques. Or pour l'instant la phase des SNP hétérozygotes que nous avons détectés n'est pas connue.

La reconstruction de la phase est possible avec des méthodes spécifiques de séquençage et d'assemblage qui permettent de distinguer les deux haplotypes. L'électrophorèse capillaire est utilisée dans ce sens (Szantai et al., 2005), tout comme la nano-PCR (Pan et al., 2012) qui permet l'amplification spécifique de chaque séquence ou encore les nanopores (Mirsaidov et al., 2010) qui détectent les séquences une par une grâce à leurs signaux électriques. Ces méthodes sont fiables car elles permettent une lecture directe des haplotypes, mais elles sont coûteuses et longues à mettre en place. Par ailleurs, des algorithmes permettant de reconstruire la phase à partir des données génomiques sans séquençage spécifique, et qui sont donc moins coûteux, ont été développés (Niu, 2004). En particulier, des programmes dédiés à la reconstruction des haplotypes à partir de données populationnelles existent, tels que PHASE (Scheet and Stephens, 2006) même si son utilisation présente le risque de bruiter le signal à cause de l'histoire démographique forte des isolats de *M. larici-populina*.

2.3. Validation des candidats, approche biochimique et moléculaire

Une fois les gènes candidats identifiés, les méthodes de génétique des populations ne permettent pas la validation des fonctions sous-jacentes et leur implication dans les phénomènes adaptatifs considérés (Bonin, 2008). Cette validation passe par des études moléculaires et biochimiques des gènes et des protéines qu'ils codent. Dans le cadre des interactions hôte-pathogène, de nombreuses études s'intéressent au dialogue moléculaire entre l'hôte et le pathogène, depuis la recherche d'homologie de séquences jusqu'à la localisation de la protéine exprimée (Petre and Kamoun, 2014).

Une des approches employées consiste en l'expression des gènes candidats dans un système dont on contrôle les conditions (hôte sensible, mutant, etc.). Pour les espèces qui ne sont pas propices à la transformation stable, comme c'est le cas pour les espèces de peuplier, on peut obtenir une transformation transitoire (agro-infiltration) ou utiliser un système hétérologue chez une espèce modèle comme *Nicotiana benthamiana*. Dans le cas de *M. larici-populina*, les tentatives d'agro-infiltration d'effecteurs candidats (initialement choisis sur la base des patrons d'expression de gènes codant des petites protéines sécrétées au cours de la phase infectieuse ; Hacquard et al., 2012) sur le peuplier se sont avérées infructueuses. Dans une étude récente, 20 effecteurs candidats de *M. larici-populina* ont été testés en système hétérologue. Cette étude a démontré que six de ces effecteurs candidats avaient une localisation cellulaire spécifique dans le nucléole, les mitochondries ou les chloroplastes, indiquant les sites possibles des interactions moléculaires de chacune de ces protéines

(Petre et al., 2015). Cette approche en système hétérologue permet aussi d'identifier des interactions potentielles de l'effecteur avec des protéines de la plante par co-immunoprécipitation suivie de l'identification des peptides sélectionnés par spectrométrie de masse. Ainsi, plusieurs effecteurs candidats exprimés chez *N. benthamiana* ont montré des interactions fortes avec des protéines candidates de plante (Petre et al., 2015). Toutefois, ces études ne sont pas encore suffisante pour démontrer le statut de gène d'avirulence de tels gènes candidats. Une autre approche envisageable, et actuellement testée au sein de l'UMR Interactions Arbres/Microorganismes, est de produire des protéines recombinantes des effecteurs candidats chez la bactérie *Escherichia coli*. Une fois produites, ces protéines peuvent être utilisées à diverses fins. Si leur infiltration dans les tissus foliaires de cultivars de peupliers porteur de résistances déployées en plantation déclenche une réaction d'hypersensibilité (HR), cela démontrerait leur statut de facteur d'avirulence. Par ailleurs, les protéines recombinantes peuvent permettre de générer des anticorps spécifiques pour réaliser l'immunolocalisation *in planta* chez l'hôte peuplier sur des coupes de feuilles infectées par *M. larici-populina* et vérifier ainsi les localisations observées en système hétérologue. Enfin, la structure des effecteurs candidats peut être déterminée par cristallographie ou par résonance magnétique nucléaire. Par ailleurs, pour formellement démontrer l'interaction R-Avr entre les produits des gènes de résistance et d'avirulence chez l'hôte et le champignon, respectivement, il faut identifier la protéine de résistance de la plante. Une telle interaction a pu être démontrée chez la rouille modèle *M. lini*, l'agent de la rouille du lin, avec notamment la protéine AvrL567 qui interagit avec les protéines de résistance L5, L6 et L7 du lin (Dodds et al., 2006). Dans le pathosystème peuplier-rouille du peuplier, il serait également envisageable d'isoler les protéines de résistance de cultivars de peuplier en passant des extraits protéiques de feuilles sur des colonnes d'affinité porteuses d'une protéine d'avirulence candidate (par exemple Avr7). Une fois les couples protéine d'avirulence/protéine de résistance identifiés, la démonstration formelle de l'interaction moléculaire pourrait être testée par des techniques telles que le système double hybride en levure, par BiFC (fluorescence par complémentation bi-moléculaire ; Kerppola, 2006) ou encore par co-immunoprécipitation. De telles interactions ont pu être montrées chez *Cladosporium fulvum* (Stergiopoulos and de Wit, 2009), *Verticillium* spp. (Thomma et al., 2011), et *M. lini* (Ellis et al., 2007).

Dans de nombreux systèmes, ces approches moléculaires et biochimiques sont conduites sur des gènes d'effecteurs candidats sélectionnés a priori selon des critères prédéfinis tels que la présence d'un peptide signal indiquant une sécrétion, la taille des protéines (les effecteurs sont le plus communément de petite taille), une richesse en cystéines, bien que ces différents critères ne constituent pas une définition absolue des effecteurs (Petre et al., 2014). De plus, les approches moléculaires et biochimiques sont longues à mettre en place dans des pathosystèmes complexes non

modèles tels que celui associant peuplier et rouille du peuplier. Ainsi, elles ne permettent, pour le moment, que d'étudier un faible nombre de candidats simultanément. Un des intérêts des études de génomique des populations est de permettre l'identification des gènes candidats sans aucun apriori de fonction ou d'homologie avec des effecteurs connus. Cette approche populationnelle devrait permettre à terme d'identifier de nouvelles classes de gènes d'avirulence et de les caractériser.

3. Et après ? Vers une gestion durable des peupliers ?

L'efficacité du déploiements des cultivars porteurs de résistances qualitatives à la rouille foliaire est remise en cause par observation des contournements systématiques de ces résistances par l'agent de la rouille du peuplier *M. larici-populina*. C'est pourquoi un certain nombre de programmes de sélection de cultivars porteurs de résistances quantitatives a été engagé. Mais, bien que plus rares, les cas de contournement de résistance quantitative existent et ce dans plusieurs pathosystèmes (Andrivon et al., 2007; Caffier et al., 2014; Delmotte et al., 2014), incluant le système peuplier-*M. Larici-populina* (Dowkiw et al., 2010). Dans le cas de *M. larici-populina*, Michaël Pernaci a démontré au cours de sa thèse une évolution rapide des traits d'agressivité des spores tel que le taux de sporulation et la taille des lésions (Pernaci, 2015). Or il a été montré que les traits d'agressivité peuvent intervenir dans le contournement de résistances quantitatives déployées au champ. Particulièrement, Van den Berg et al., (2014) ont montré que le taux de sporulation du champignon peut augmenter, permettant le contournement d'un gène de résistance quantitatif. Ainsi, le déploiement de résistances quantitatives semble plus adapté que celui de résistances qualitatives mais il nécessite des études préalables sur le potentiel adaptatif du pathogène de manière à éviter ainsi de tels contournements.

Chez *M. larici-populina*, une carte génétique a pu être construite à partir du séquençage de 95 descendants issus de l'autofécondation de l'isolat de référence (98AG31) et de la détection de points de recombinaison. Cette étude a permis d'identifier les régions génomiques intervenant dans l'expression des caractères d'agressivité. Un QTL de virulence/avirulence nommé AvrX a ainsi pu être associé au contournement d'une résistance du peuplier qui n'avait pas été décrite précédemment (Pernaci, 2015). Cependant cette étude n'a pas pu permettre d'analyser le déterminisme des huit facteurs de virulence connus, aucun n'étant en ségrégation dans la descendance de l'autofécondation. Ainsi, un nouveau projet, démarrant au sein de l'équipe, consiste en l'analyse d'une nouvelle descendance issue du croisement de deux souches de *M. larici-populina* présentant des combinaisons de virulences différentes. Cette étude pourrait permettre l'identification de QTLs pour différentes virulences et permettre ainsi de progresser dans l'étude et la compréhension des mécanismes moléculaires d'interaction sous-jacent dans le pathosystème. A terme, une meilleure connaissance des mécanismes d'interaction et de l'évolution du système pourrait aider à guider les

stratégies de gestion des résistances en plantation afin de viser une résistance plus durable à la maladie. Certaines virulences sont fixées dans les populations (c'est le cas de la virulence 4), rendant leur détection impossible par une telle approche. De même, les déterminants des résistances quantitatives peuvent être difficiles à détecter même si elle sont en ségrégation, du fait du faible effet moyen des gènes.

Des travaux récents au sein du laboratoire (Hayden et al. non publié) ont démontré un coût de la virulence (augmentation du temps de latence chez les individus virulents 7 de 1994) et une spécificité des contournements de résistance, à savoir que les individus de *M. larici-populina* porteurs des différentes virulences se répartissent au niveau territorial dans les zones de plantations de peupliers porteurs des résistances correspondantes, suggérant que les gènes de virulence ne sont pas adaptatifs en l'absence de la résistance correspondante. Ces informations suggèrent que le gène de virulence ne serait pas le seul impliqué dans le contournement. Or une étude de génomique des populations, telle que celle présentée dans ce travail, permet d'identifier l'ensemble des régions génomiques sous sélection et donc potentiellement l'ensemble des gènes impliqués dans le contournement et pas seulement ceux de la virulence. Les phénotypes (à savoir les traits morphologiques et d'agressivité) des isolats de *M. larici-populina*, porteurs des différentes virulences, ont été mesurés. Les corrélations phénotypes/génotypes entre ces données et les études de scan génomique devraient permettre d'identifier les gènes impliqués dans les traits de vie de *M. larici-populina* qui ont été sélectionnés lors des différents contournements de résistances et qui présentent ainsi un fort potentiel adaptatif vis-à-vis de nouvelles résistances qui seraient déployées dans le futur. A terme, ce type de méthode devrait permettre de développer et déployer des résistances quantitatives plus durables, en évitant de déployer des résistances qualitatives ou quantitatives qui seraient faciles à contourner pour le pathogène.

BIBLIOGRAPHIE

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. (2000). The genome sequence of *Drosophila melanogaster*. Science *287*, 2185–2195.

Aime, M.C., Matheny, P.B., Henk, D.A., Frieders, E.M., Nilsson, R.H., Piepenbring, M., McLaughlin, D.J., Szabo, L.J., Begerow, D., Sampaio, J.P., et al. (2006). An overview of the higher level classification of Pucciniomycotina based on combined analyses of nuclear large and small subunit rDNA sequences. Mycologia *98*, 896-905.

Akey, J.M. (2009). Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res. *19*, 711–722.

Akiba, T., Koyama, K., Ishiki, Y., Kimura, S., and Fukushima, T. (1960). On the mechanism of the development of multiple-drug-resistant clones of *Shigella. Jpn*. J. Microbiol. *4*, 219–227.

Anderson, E.C., Williamson, E.G., and Thompson, E.A. (2000). Monte Carlo evaluation of the likelihood for N(e) from temporally spaced samples. Genetics *156*, 2109–2118.

Andrivon, D., Pilet, F., Montarry, J., Hafidi, M., Corbiere, R., Achbani, E.H., Pelle, R., and Ellisseche, D. (2007). Adaptation of *Phytophthora infestans* to Partial Resistance in Potato: Evidence from French and Moroccan Populations. Phytopathology *97*, 338–343.

Balloux, F., Lehmann, L., and de Meeus, T. (2003). The population genetics of clonal and partially clonal diploids. Genetics *164*, 1635–1644.

Balter, M. (2007). Plant science. Seeking agriculture's ancient roots. Science *316*, 1830–1835.

Bank, C., Ewing, G.B., Ferrer-Admettla, A., Foll, M., and Jensen, J.D. (2014). Thinking too positive? Revisiting current methods of population genetic selection inference. Trends Genet. TIG *30*, 540–546.

Barres, B., Halkett, F., Dutech, C., Andrieux, A., Pinon, J., and Frey, P. (2008). Genetic structure of the poplar rust fungus *Melampsora larici-populina*: evidence for isolation by distance in Europe and recent founder effects overseas. Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis. *8*, 577–587.

Baxter, L., Tripathy, S., Ishaque, N., Boot, N., Cabral, A., Kemen, E., Thines, M., Ah-Fong, A.,

Anderson, R., Badejoko, W., et al. (2010). Signatures of adaptation to obligate biotrophy in the Hyaloperonospora arabidopsidis genome. Science *330*, 1549–1551.

Bazin, E., Dawson, K.J., and Beaumont, M.A. (2010). Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. Genetics *185*, 587–602.

Beaumont, M.A., and Balding, D.J. (2004). Identifying adaptive genetic divergence among populations from genome scans. Mol. Ecol. *13*, 969–980.

Beaumont, M.A., and Rannala, B. (2004). The Bayesian revolution in genetics. Nat Rev Genet *5*, 251–261.

Begun, D.J., Holloway, A.K., Stevens, K., Hillier, L.W., Poh, Y.-P., Hahn, M.W., Nista, P.M., Jones, C.D., Kern, A.D., Dewey, C.N., et al. (2007). Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in *Drosophila simulans*. PLoS Biol *5*, e310.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53–59.

Bonin, A. (2008). Population genomics: a new generation of genome scans to bridge the gap with functional genomics. Mol Ecol *17*, 3583–3584.

Bourassa, M., Bernier, L., and Hamelin, R.C. (2007). Genetic Diversity in Poplar Leaf Rust (*Melampsora medusae* f. sp. *deltoidae*) in the Zones of Host Sympatry and Allopatry. Phytopathology *97*, 603–610.

Boutemy, L.S., King, S.R.F., Win, J., Hughes, R.K., Clarke, T.A., Blumenschein, T.M.A., Kamoun, S., and Banfield, M.J. (2011). Structures of Phytophthora RXLR effector proteins: a conserved but adaptable fold underpins functional diversity. J. Biol. Chem. *286*, 35834–35842.

Burke, M.K. (2012). How does adaptation sweep through the genome? Insights from long-term selection experiments. Proc. Biol. Sci. *279*, 5029–5038.

Cadzow, M., Boocock, J., Nguyen, H.T., Wilcox, P., Merriman, T.R., and Black, M.A. (2014). A bioinformatics workflow for detecting signatures of selection in genomic data. Front. Genet. *5*, 293.

Caffier, V., Lasserre-Zuber, P., Giraud, M., Lascostes, M., Stievenard, R., Lemarquand, A., van de Weg, E., Expert, P., Denance, C., Didelot, F., et al. (2014). Erosion of quantitative host resistance in the applex *Venturia inaequalis* pathosystem. Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet.

Infect. Dis. *27*, 481–489.

Cantu, D., Govindarajulu, M., Kozik, A., Wang, M., Chen, X., Kojima, K.K., Jurka, J., Michelmore, R.W., and Dubcovsky, J. (2011). Next generation sequencing provides rapid access to the genome of *Puccinia striiformis* f. sp. *tritici*, the causal agent of wheat stripe rust. PloS One *6*, e24230.

Cantu, D., Segovia, V., MacLean, D., Bayles, R., Chen, X., Kamoun, S., Dubcovsky, J., Saunders, D.G., and Uauy, C. (2013). Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f. sp. *tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. BMC Genomics *14*, 270.

Catanzariti, A.M., Dodds, P.N., and Ellis, J.G. (2007). Avirulence proteins from haustoria-forming pathogens. FEMS Microbiol Lett *269*, 181–188.

C. elegans Sequencing Consortium. (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science *282*, 2012–2018.

Charlesworth, B. (2013). Background selection 20 years on: the Wilhelmine E. Key 2012 invitational lecture. J. Hered. *104*, 161–171.

Cornuet, J.M., and Luikart, G. (1996). Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics *144*, 2001–2014.

Dai, Y., Jia, Y., Correll, J., Wang, X., and Wang, Y. (2010). Diversification and evolution of the avirulence gene AVR-Pita1 in field isolates of *Magnaporthe oryzae*. Fungal Genet. Biol. FG B *47*, 973–980.

Danchin, E.G.J., Rosso, M.-N., Vieira, P., de Almeida-Engler, J., Coutinho, P.M., Henrissat, B., and Abad, P. (2010). Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. Proc. Natl. Acad. Sci. U. S. A. *107*, 17651–17656.

Dangl, J.L., Horvath, D.M., and Staskawicz, B.J. (2013). Pivoting the plant immune system from dissection to deployment. Science *341*, 746–751.

Das, S., and Vikalo, H. (2013). Base calling for high-throughput short-read sequencing: dynamic programming solutions. BMC Bioinformatics *14*, 129.

Delmotte, F., Mestre, P., Schneider, C., Kassemeyer, H.-H., Kozma, P., Richart-Cervera, S., Rouxel, M., and Deliere, L. (2014). Rapid and multiregional adaptation to host partial resistance in a plant pathogenic oomycete: evidence from European populations of *Plasmopara viticola*, the causal agent

of grapevine downy mildew. Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis. *27*, 500–508.

De Mita, S., and Siol, M. (2012). EggLib: processing, analysis and simulation tools for population genetics and genomics. BMC Genet *13*, 27.

De Mita, S., Ronfort, J., McKhann, H.I., Poncet, C., El Malki, R., and Bataillon, T. (2007). Investigation of the demographic and selective forces shaping the nucleotide diversity of genes involved in nod factor signaling in *Medicago truncatula*. Genetics *177*, 2123–2133.

Demogines, A., Wong, A., Aquadro, C., and Alani, E. (2008). Incompatibilities involving yeast mismatch repair genes: a role for genetic modifiers and implications for disease penetrance and variation in genomic mutation rates. PLoS Genet. *4*, e1000103.

Depaulis, F., Mousset, S., and Veuille, M. (2003). Power of neutrality tests to detect bottlenecks and hitchhiking. J Mol Evol *57 Suppl 1*, S190–S200.

Devault, A.M., McLoughlin, K., Jaing, C., Gardner, S., Porter, T.M., Enk, J.M., Thissen, J., Allen, J., Borucki, M., DeWitte, S.N., et al. (2014). Ancient pathogen DNA in archaeological samples detected with a Microbial Detection Array. Sci. Rep. *4*, 4245.

Dodds, P.N., and Rathjen, J.P. (2010). Plant immunity: towards an integrated view of plant-pathogen interactions. Nat. Rev. Genet. *11*, 539–548.

Dodds, P.N., Lawrence, G.J., Catanzariti, A.-M., Teh, T., Wang, C.-I.A., Ayliffe, M.A., Kobe, B., and Ellis, J.G. (2006). Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. Proc. Natl. Acad. Sci. U. S. A. *103*, 8888–8893.

Dowkiw, A., Voisin, E., and Bastien, C. (2010). Potential of Eurasian poplar rust to overcome a major quantitative resistance factor. Plant Pathol. *59*, 523–534.

Drummond, A.J., and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. *7*, 214.

Duplessis, S., Cuomo, C.A., Lin, Y.-C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., Joly, D.L., Hacquard, S., Amselem, J., Cantarel, B.L., et al. (2011a). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. Proc. Natl. Acad. Sci. U. S. A. *108*, 9166–9171.

Duplessis, S., Hacquard, S., Delaruelle, C., Tisserant, E., Frey, P., Martin, F., and Kohler, A. (2011b).

Melampsora larici-populina transcript profiling during germination and timecourse infection of poplar leaves reveals dynamic expression patterns associated with virulence and biotrophy. Mol Plant Microbe Interact *24*, 808–818.

Duplessis, S., Spanu, P.D., and Schirawski, J. (2013). Biotrophic Fungi (Powdery Mildews, Rusts, and Smuts). In The Ecological Genomics of Fungi, (John Wiley & Sons, Inc), pp. 149–168.

Duplessis, S., Bakkeren, G., and Hamelin, R. (2014). Chapter Six - Advancing Knowledge on Biology of Rust Fungi Through Genomics. In Advances in Botanical Research, Francis M. Martin, ed. (Academic Press), pp. 173–209.

Durand, E., Jay, F., Gaggiotti, O.E., and Francois, O. (2009). Spatial inference of admixture proportions and secondary contact zones. Mol. Biol. Evol. *26*, 1963–1973.

Ellis, J.G., Dodds, P.N., and Lawrence, G.J. (2007). Flax rust resistance gene specificity is based on direct resistance-avirulence protein interactions. Annu. Rev. Phytopathol. *45*, 289-306.

Ellison, C.E., Hall, C., Kowbel, D., Welch, J., Brem, R.B., Glass, N.L., and Taylor, J.W. (2011). Population genomics and local adaptation in wild isolates of a model microbial eukaryote. Proc Natl Acad Sci U A *108*, 2831–2836.

Erickson, R.P. (2003). Somatic gene mutation and human disease other than cancer. Mutat. Res. *543*, 125–136.

Fay, J.C., and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. Genetics *155*, 1405–1413.

Flor, H.H. (1971). Current Status of the Gene-For-Gene Concept. Annu. Rev. Phytopathol. *9*, 275–296.

Freeman, B.C., and Beattie, G.A. (2008). An Overview of Plant Defenses against Pathogens and Herbivores. Plant Health Instr.

Fu, Y.X., and Li, W.H. (1993). Statistical tests of neutrality of mutations. Genetics *133*, 693-709.

Fudal, I., Ross, S., Brun, H., Besnard, A.-L., Ermel, M., Kuhn, M.-L., Balesdent, M.-H., and Rouxel, T. (2009). Repeat-induced point mutation (RIP) as an alternative mechanism of evolution toward virulence in *Leptosphaeria maculans*. Mol. Plant-Microbe Interact. *22*, 932–941.

Gerard, P.R., Husson, C., Pinon, J., and Frey, P. (2006). Comparison of Genetic and Virulence

Diversity of *Melampsora larici-populina* Populations on Wild and Cultivated Poplar and Influence of the Alternate Host. Phytopathology *96*, 1027–1036.

Gilles, A., Meglecz, E., Pech, N., Ferreira, S., Malausa, T., and Martin, J.-F. (2011). Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics *12*, 245.

Gladieux, P., Zhang, X.G., Roldan-Ruiz, I., Caffier, V., Leroy, T., Devaux, M., Van Glabeke, S., Coart, E., and Le Cam, B. (2010). Evolution of the population structure of *Venturia inaequalis*, the apple scab fungus, associated with the domestication of its host. Mol Ecol *19*, 658–674.

Glenn, T.C. (2011). Field guide to next-generation DNA sequencers. Mol Ecol Resour *11*, 759–769.

Goff, S.A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science *296*, 92-100.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. Science *274*, 546, 563–567.

Gout, L., Fudal, I., Kuhn, M.L., Blaise, F., Eckert, M., Cattolico, L., Balesdent, M.H., and Rouxel, T. (2006). Lost in the middle of nowhere: the AvrLm1 avirulence gene of the Dothideomycete *Leptosphaeria maculans*. Mol Microbiol *60*, 67–80.

Gout, L., Kuhn, M.L., Vincenot, L., Bernard-Samain, S., Cattolico, L., Barbetti, M., Moreno-Rico, O., Balesdent, M.-H., and Rouxel, T. (2007). Genome structure impacts molecular evolution at the AvrLm1 avirulence locus of the plant pathogen *Leptosphaeria maculans*. Environ. Microbiol. *9*, 2978–2992.

Goyeau, H., Halkett, F., Zapater, M.-F., Carlier, J., and Lannou, C. (2007). Clonality and host selection in the wheat pathogenic fungus *Puccinia triticina*. Fungal Genet. Biol. FG B *44*, 474–483.

Graffelman, J., Sánchez, M., Cook, S., and Moreno, V. (2013). Statistical Inference for Hardy-Weinberg Proportions in the Presence of Missing Genotype Information. PLoS ONE *8*, e83316.

Gregory, S.G., Barlow, K.F., McLay, K.E., Kaul, R., Swarbreck, D., Dunham, A., Scott, C.E., Howe, K.L., Woodfine, K., Spencer, C.C.A., et al. (2006). The DNA sequence and biological annotation of human chromosome 1. Nature *441*, 315–321.

Guerin, F., Gladieux, P., and Le Cam, B. (2007). Origin and colonization history of newly virulent strains of the phytopathogenic fungus *Venturia inaequalis*. Fungal Genet Biol *44*, 284–292.

Haas, B.J., Kamoun, S., Zody, M.C., Jiang, R.H.Y., Handsaker, R.E., Cano, L.M., Grabherr, M., Kodira, C.D., Raffaele, S., Torto-Alalibo, T., et al. (2009). Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. Nature *461*, 393–398.

Hacquard, S. (2010). Contribution à l'étude des déterminants génétiques impliqués dans le processus infectieux de *Melampsora larici-populina*, l'agent de la rouille foliaire du peuplier.

Hacquard, S., Petre, B., Frey, P., Hecker, A., Rouhier, N., and Duplessis, S. (2011). The poplar-poplar rust interaction: insights from genomics and transcriptomics. J Pathog *2011*, 716041.

Hacquard, S., Joly, D.L., Lin, Y.-C., Tisserant, E., Feau, N., Delaruelle, C., Legue, V., Kohler, A., Tanguay, P., Petre, B., et al. (2012). A comprehensive analysis of genes encoding small secreted proteins identifies candidate effectors in *Melampsora larici-populina* (poplar leaf rust). Mol. Plant-Microbe Interact. *25*, 279–293.

Hacquard, S., Kracher, B., Maekawa, T., Vernaldi, S., Schulze-Lefert, P., and Ver Loren van Themaat, E. (2013). Mosaic genome structure of the barley powdery mildew pathogen and conservation of transcriptional programs in divergent hosts. Proc. Natl. Acad. Sci. U. S. A. *110*, E2219–E2228.

Halkett, F., Simon, J.C., and Balloux, F. (2005). Tackling the population genetics of clonal and partially clonal organisms. Trends Ecol Evol *20*, 194–201.

Haudry, A., Cenci, A., Ravel, C., Bataillon, T., Brunel, D., Poncet, C., Hochu, I., Poirier, S., Santoni, S., Glemin, S., et al. (2007). Grinding up wheat: a massive loss of nucleotide diversity since domestication. Mol. Biol. Evol. *24*, 1506–1517.

Hermisson, J., and Pennings, P.S. (2005). Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics *169*, 2335–2352.

Hudson, R.R., Kreitman, M., and Aguade, M. (1987). A test of neutral molecular evolution based on nucleotide data. Genetics *116*, 153–159.

Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., and Welch, D.M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. *8*, R143.

Jin, Y., Szabo, L.J., and Carson, M. (2010). Century-old mystery of Puccinia striiformis life history solved with the identification of Berberis as an alternate host. Phytopathology *100*, 432–435.

Jones, J.D., and Dangl, J.L. (2006). The plant immune system. Nature *444*, 323–329.

Joosten, M.H., Cozijnsen, T.J., and De Wit, P.J. (1994). Host resistance to a fungal tomato pathogen lost by a single base-pair change in an avirulence gene. Nature *367*, 384–386.

Kaplan, N.L., Hudson, R.R., and Langley, C.H. (1989). The "hitchhiking effect" revisited. Genetics *123*, 887–899.

Kerppola, T.K. (2006). Visualization of molecular interactions by fluorescence complementation. Nat. Rev. Mol. Cell Biol. *7*, 449–456.

Khang, C.H., Park, S.Y., Lee, Y.H., Valent, B., and Kang, S. (2008). Genome organization and evolution of the AVR-Pita avirulence gene family in the *Magnaporthe grisea* species complex. Mol Plant Microbe Interact *21*, 658–670.

Kim, Y., and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics *160*, 765–777.

Kimura, R., Fujimoto, A., Tokunaga, K., and Ohashi, J. (2007). A practical genome scan for population-specific strong selective sweeps that have reached fixation. PLoS One *2*, e286.

Kingman, J.F. (2000). Origins of the coalescent. 1974-1982. Genetics *156*, 1461–1463.

Kircher, M., Stenzel, U., and Kelso, J. (2009). Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biol. *10*, R83.

Knief, C. (2014). Analysis of plant microbe interactions in the era of next generation sequencing technologies. Front. Plant Sci. *5*, 216.

Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D., et al. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat. Biotechnol. *30*, 693-700.

Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M., and Turner, D.J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat. Methods *6*, 291–295.

Laurans, F., and Pilate, G. (1999). Histological Aspects of a Hypersensitive Response in Poplar to *Melampsora larici-populina*. Phytopathology *89*, 233–238.

Li, H., Zhang, Y., Zhang, Y.-P., and Fu, Y.-X. (2003). Neutrality tests using DNA polymorphism from multiple samples. Genetics *163*, 1147–1151.

Li, M., Tian, S., Yeung, C.K.L., Meng, X., Tang, Q., Niu, L., Wang, X., Jin, L., Ma, J., Long, K., et al. (2014). Whole-genome sequencing of Berkshire (European native pig) provides insights into its origin and domestication. Sci. Rep. *4*, 4678.

Liti, G., Carter, D.M., Moses, A.M., Warringer, J., Parts, L., James, S.A., Davey, R.P., Roberts, I.N., Burt, A., Koufopanou, V., et al. (2009). Population genomics of domestic and wild yeasts. Nature *458*, 337–341.

Liu, B., Li, J.-F., Ao, Y., Qu, J., Li, Z., Su, J., Zhang, Y., Liu, J., Feng, D., Qi, K., et al. (2012). Lysin motif-containing proteins LYP4 and LYP6 play dual roles in peptidoglycan and chitin perception in rice innate immunity. Plant Cell *24*, 3406–3419.

Loehrer, M., Vogel, A., Huettel, B., Reinhardt, R., Benes, V., Duplessis, S., Usadel, B., and Schaffrath, U. (2014). On the current status of *Phakopsora pachyrhizi* genome sequencing. Front. Plant Sci. *5*, 377.

Lombaert, E., Guillemaud, T., Lundgren, J., Koch, R., Facon, B., Grez, A., Loomans, A., Malausa, T., Nedved, O., Rhule, E., et al. (2014). Complementarity of statistical treatments to reconstruct worldwide routes of invasion: the case of the Asian ladybird Harmonia axyridis. Mol. Ecol. *23*, 5979–5997.

Lo Presti, L., Lanver, D., Schweizer, G., Tanaka, S., Liang, L., Tollot, M., Zuccaro, A., Reissmann, S., and Kahmann, R. (2015). Fungal effectors and plant susceptibility. Annu. Rev. Plant Biol. *66*, 513–545.

Luikart, G., England, P.R., Tallmon, D., Jordan, S., and Taberlet, P. (2003). The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet *4*, 981–994.

Luo, C., Tsementzi, D., Kyrpides, N., Read, T., and Konstantinidis, K.T. (2012a). Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PloS One *7*, :e30087.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012b). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience *1*, 18.

Macho, A.P., and Zipfel, C. (2014). Plant PRRs and the activation of innate immune signaling. Mol. Cell *54*, 263–272.

Mardis, E.R. (2008). Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet *9*, 387–402.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376–380.

McDonald, B.A. (2004). Population Genetics of Plant Pathogens. Plant Health Instr.

McDonald, J.H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature *351*, 652–654.

McDowell, J.M. (2011). Genomes of obligate plant pathogens reveal adaptations for obligate parasitism. Proc. Natl. Acad. Sci. U. S. A. *108*, 8921–8922.

McPherson, J.D., Marra, M., Hillier, L., Waterston, R.H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E.R., Wilson, R.K., et al. (2001). A physical map of the human genome. Nature *409*, 934–941.

McTaggart, A.R., Shivas, R.G., van der Nest, M.A., Roux, J., Wingfield, B.D., and Wingfield, M.J. (2015). Host jumps shaped the diversity of extant rust fungi (Pucciniales). New Phytol.

Mirsaidov, U.M., Wang, D., Timp, W., and Timp, G. (2010). Molecular diagnostics for personal medicine using a nanopore. Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol. *2*, 367–381.

Miya, A., Albert, P., Shinya, T., Desaki, Y., Ichimura, K., Shirasu, K., Narusaka, Y., Kawakami, N., Kaku, H., and Shibuya, N. (2007). CERK1, a LysM receptor kinase, is essential for chitin elicitor signaling in Arabidopsis. Proc. Natl. Acad. Sci. U. S. A. *104*, 19613–19618.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M.C., Hirai, A., Takahashi, H., et al. (2011). Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. *39*, e90.

Namroud, M.C., Beaulieu, J., Juge, N., Laroche, J., and Bousquet, J. (2008). Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. Mol Ecol *17*, 3599–3613.

Nemri, A., Saunders, D.G.O., Anderson, C., Upadhyaya, N.M., Win, J., Lawrence, G.J., Jones, D.A., Kamoun, S., Ellis, J.G., and Dodds, P.N. (2014). The genome sequence and effector complement of the flax rust pathogen *Melampsora lini*. Front. Plant Sci. *5*, 98.

Nguyen, P., Ma, J., Pei, D., Obert, C., Cheng, C., and Geiger, T.L. (2011). Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. BMC Genomics *12*, 106.

Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J., et al. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol *3*, e170.

Niu, T. (2004). Algorithms for inferring haplotypes. Genet. Epidemiol. *27*, 334–347.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. Nature *246*, 96–98.

Ohta, and Gillespie (1996). Development of Neutral and Nearly Neutral Theories. Theor. Popul. Biol. *49*, 128–142.

Oleksyk, T.K., Smith, M.W., and O'Brien, S.J. (2010). Genome-wide scans for footprints of natural selection. Philos Trans R Soc Lond B Biol Sci *365*, 185–205.

Orbach, M.J., Farrall, L., Sweigard, J.A., Chumley, F.G., and Valent, B. (2000). A telomeric avirulence gene determines efficacy for the rice blast resistance gene Pi-ta. Plant Cell *12*, 2019–2032.

Pan, D., Mi, L., Huang, Q., Hu, J., and Fan, C. (2012). Genetic analysis with nanoPCR. Integr. Biol. Quant. Biosci. Nano Macro *4*, 1155–1163.

Pariset, L., Joost, S., Marsan, P.A., and Valentini, A. (2009). Landscape genomics and biased FST approaches reveal single nucleotide polymorphisms under selection in goat breeds of North-East Mediterranean. BMC Genet *10*, 7.

Pernaci, M. (2015). Étude des traits d'histoire de vie de "*Melampsora larici-populina*", agent de la rouille du peuplier : de leur déterminisme génétique à leurs conséquences évolutives. Thèse de doctorat en Biologie végétale et forestière. Université de Lorraine.

Persoons, A., Morin, E., Delaruelle, C., Payen, T., Halkett, F., Frey, P., De Mita, S., and Duplessis, S. (2014). Patterns of genomic variation in the poplar rust fungus *Melampsora larici-populina* identify pathogenesis-related factors. Front. Plant Sci. *5*, 540.

Petre, B., and Kamoun, S. (2014). How do filamentous pathogens deliver effector proteins into plant cells? PLoS Biol. *12*, :e1001801.

Petre, B., Joly, D.L., and Duplessis, S. (2014). Effector proteins of rust fungi. Front. Plant Sci. *5*, 416.

Petre, B., Saunders, D.G.O., Sklenar, J., Lorrain, C., Win, J., Duplessis, S., and Kamoun, S. (2015). Candidate effector proteins of the rust pathogen *Melampsora larici-populina* target diverse plant cell compartments. Mol. Plant-Microbe Interact. MPMI *28*, 689-700.

Pinon, J., and Frey, P. (1997). Structure of *Melampsora larici-populina* populations on wild and cultivated poplar. Eur. J. Plant Pathol. *103*, 159–173.

Pinon, J. and Frey, P. (2005). Interactions between poplar clones and *Melampsora* populations and their implications for breeding for durable resistance. In: Rust Dis. Willow Poplar. Pei, M. H., McCracken, A. R., 139–154

van Poppel, P.M., Guo, J., van de Vondervoort, P.J., Jung, M.W., Birch, P.R., Whisson, S.C., and Govers, F. (2008). The *Phytophthora infestans* avirulence gene Avr4 encodes an RXLR-dEER effector. Mol Plant Microbe Interact *21*, 1460–1470.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics *13*, 241.

Quillery, E., Quenez, O., Peterlongo, P., and Plantard, O. (2014). Development of genomic resources for the tick Ixodes ricinus: isolation and characterization of single nucleotide polymorphisms. Mol. Ecol. Resour. *14(*2), 393-400.

Raffaele, S., and Kamoun, S. (2012). Genome evolution in filamentous plant pathogens: why bigger can be better. Nat. Rev. Microbiol. *10*, 417–430.

Raffaele, S., Win, J., Cano, L.M., and Kamoun, S. (2010). Analyses of genome architecture and gene expression reveal novel candidate virulence factors in the secretome of *Phytophthora infestans*. BMC Genomics *16*, 11:637.

Rafiqi, M., Gan, P.H., Ravensdale, M., Lawrence, G.J., Ellis, J.G., Jones, D.A., Hardham, A.R., and Dodds, P.N. (2010). Internalization of flax rust avirulence proteins into flax and tobacco cells can occur in the absence of the pathogen. Plant Cell *22*, 2017–2032.

Rafiqi, M., Ellis, J.G., Ludowici, V.A., Hardham, A.R., and Dodds, P.N. (2012). Challenges and progress towards understanding the role of effectors in plant-fungal interactions. Curr Opin Plant Biol *15*, 477–482.

Rinaldi, C., Kohler, A., Frey, P., Duchaussoy, F., Ningre, N., Couloux, A., Wincker, P., Le Thiec, D., Fluch, S., Martin, F., et al. (2007). Transcript profiling of poplar leaves upon infection with compatible and incompatible strains of the foliar rust *Melampsora larici-populina*. Plant. Physiol. *144*, 347–366.

Rogozin, I.B., and Pavlov, Y.I. (2003). Theoretical analysis of mutation hotspots and their DNA sequence context specificity. Mutat. Res. *544*(1), 65-85.

Rouxel, T., and Balesdent, M.H. (2005). The stem canker (blackleg) fungus, Leptosphaeria maculans, enters the genomic era. Mol. Plant Pathol. *6*, 225–241.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature *419*, 832–837.

Saunders, D.G., Win, J., Cano, L.M., Szabo, L.J., Kamoun, S., and Raffaele, S. (2012). Using hierarchical clustering of secreted protein families to classify and rank candidate effectors of rust fungi. PLoS One *7*, e29847.

Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. *78*, 629–644.

Schneider, D.J., and Collmer, A. (2010). Studying plant-pathogen interactions in the genomics era: beyond molecular Koch's postulates to systems biology. Annu. Rev. Phytopathol. *48*, 457–479.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nat. Biotechnol. *26*, 1135–1145.

Siol, M., Wright, S.I., and Barrett, S.C. (2010). The population genomics of plant adaptation. New Phytol *188*, 313–332.

Soanes, D., and Richards, T.A. (2014). Horizontal gene transfer in eukaryotic plant pathogens. Annu. Rev. Phytopathol. *52*, 583-614..

Spanu, P.D., Abbott, J.C., Amselem, J., Burgis, T.A., Soanes, D.M., Stuber, K., Ver Loren van Themaat, E., Brown, J.K.M., Butcher, S.A., Gurr, S.J., et al. (2010). Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. Science *330*, 1543–1546.

Stahl, E.A., Dwyer, G., Mauricio, R., Kreitman, M., and Bergelson, J. (1999). Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. Nature *400*, 667–671.

Stergiopoulos, I., and de Wit, P.J. (2009). Fungal effector proteins. Annu. Rev. Phytopathol. *47*, 233–263.

Stukenbrock, E.H., and Bataillon, T. (2012). A population genomics perspective on the emergence and adaptation of new plant pathogens in agro-ecosystems. PLoS Pathog. *8*(9),e1002893..

Stukenbrock, E.H., and McDonald, B.A. (2008). The origins of plant pathogens in agro-ecosystems. Annu. Rev. Phytopathol. *46*, 75–100.

Stukenbrock, E.H., Bataillon, T., Dutheil, J.Y., Hansen, T.T., Li, R., Zala, M., McDonald, B.A., Wang, J., and Schierup, M.H. (2011). The making of a new pathogen: insights from comparative population genomics of the domesticated wheat pathogen *Mycosphaerella graminicola* and its wild sister species. Genome Res *21*, 2157–2166.

Szantai, E., Ronai, Z., Szilagyi, A., Sasvari-Szekely, M., and Guttman, A. (2005). Haplotyping by capillary electrophoresis. J. Chromatogr. A *1079*, 41–49.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. Genetics *105*, 437–460.

Tang, K., Thornton, K.R., and Stoneking, M. (2007). A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol. *5*, e171.

Taylor, J.W., Jacobson, D.J., and Fisher, M.C. (1999). The evolution of asexual fungi: Reproduction, Speciation and Classification. Annu. Rev. Phytopathol. *37*,197-246.

Terauchi, R., and Yoshida, K. (2010). Towards population genomics of effector-effector target interactions. New Phytol. *187*, 929–939.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature *408*.

Thomma, B.P.H.J., Nurnberger, T., and Joosten, M.H.A.J. (2011). Of PAMPs and effectors: the blurred PTI-ETI dichotomy. Plant Cell *23(1)*,4-15.

Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., et al. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science *313*, 1596–1604.

Tyler, B.M., Tripathy, S., Zhang, X., Dehal, P., Jiang, R.H.Y., Aerts, A., Arredondo, F.D., Baxter, L.,

Bensasson, D., Beynon, J.L., et al. (2006). Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science *313*, 1261–1266.

Uricaru, R., Rizk, G., Lacroix, V., Quillery, E., Plantard, O., Chikhi, R., Lemaitre, C., and Peterlongo, P. (2015). Reference-free detection of isolated SNPs. Nucleic Acids Res. *43*(2),e11.

Van den Berg, F., Lannou, C., Gilligan, C.A., and van de Bosch, F. (2014). Quantitative resistance can lead to evolutionary changes in traits not targeted by the resistance QTLs. Evol. Appl. *7*, 370–380.

Van de Wouw, A.P., Cozijnsen, A.J., Hane, J.K., Brunner, P.C., McDonald, B.A., Oliver, R.P., and Howlett, B.J. (2010). Evolution of linked avirulence effectors in Leptosphaeria maculans is affected by genomic environment and exposure to resistance genes in host plants. PLoS Pathog *6(11)*,e100118.

Ve, T., Williams, S.J., Catanzariti, A.-M., Rafiqi, M., Rahman, M., Ellis, J.G., Hardham, A.R., Jones, D.A., Anderson, P.A., Dodds, P.N., et al. (2013). Structures of the flax-rust effector AvrM reveal insights into the molecular basis of plant-cell entry and effector-triggered immunity. Proc. Natl. Acad. Sci. U. S. A. *110*, 17594–17599.

Veeramah, K.R., Wegmann, D., Woerner, A., Mendez, F.L., Watkins, J.C., Destro-Bisol, G., Soodyall, H., Louie, L., and Hammer, M.F. (2012). An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. Mol. Biol. Evol. *29*, 617–630.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351.

Vialle, A., Frey, P., Hambleton, S., Bernier, L., and Hamelin, R. (2011). Poplar rust systematics and refinement of Melampsora species delineation. Fungal Divers. *50*, 227–248.

Wall, J.D. (1999). Recombination and the power of statistical tests of neutrality. Genet. Res. *74*, 65–79.

Wan, J., Zhang, X.-C., Neece, D., Ramonell, K.M., Clough, S., Kim, S.-Y., Stacey, M.G., and Stacey, G. (2008). A LysM receptor-like kinase plays a critical role in chitin signaling and fungal resistance in Arabidopsis. Plant Cell *20*, 471–481.

Wang, M., Yu, Y., Haberer, G., Marri, P.R., Fan, C., Goicoechea, J.L., Zuccolo, A., Song, X., Kudrna,

D., Ammiraju, J.S.S., et al. (2014). The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. Nat. Genet. *46*, 982–988.

Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. *7 :* 256–276.

Whitlock, M.C., and McCauley, D.E. (1999). Indirect measures of gene flow and migration: FST not equal to 1/(4Nm + 1). Hered. *82:* 117–125.

Wicker, T., Oberhaensli, S., Parlange, F., Buchmann, J.P., Shatalina, M., Roffler, S., Ben-David, R., Dolezel, J., Simkova, H., Schulze-Lefert, P., et al. (2013). The wheat powdery mildew genome shows the unique evolution of an obligate biotroph. Nat. Genet. *45*, 1092–1096.

Wright, S. (1931). Evolution in Mendelian Populations. Genetics *16*(2):97-159.

Xhaard, C., Andrieux, A., Halkett, F., and Frey, P. (2009). Characterization of 41 microsatellite loci developed from the genome sequence of the poplar rust fungus, *Melampsora larici-populina*. Conserv. Genet. Resour. *1*, 21–25.

Xhaard, C., Fabre, B., Andrieux, A., Gladieux, P., Barres, B., Frey, P., and Halkett, F. (2011). The genetic structure of the plant pathogenic fungus *Melampsora larici-populina* on its wild host is extensively impacted by host domestication. Mol Ecol *20*, 2739–2755.

Xu, X., Liu, X., Ge, S., Jensen, J.D., Hu, F., Li, X., Dong, Y., Gutenkunst, R.N., Fang, L., Huang, L., et al. (2012). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat. Biotechnol. *30*, 105–111.

Yaeno, T., Li, H., Chaparro-Garcia, A., Schornack, S., Koshiba, S., Watanabe, S., Kigawa, T., Kamoun, S., and Shirasu, K. (2011). Phosphatidylinositol monophosphate-binding interface in the oomycete RXLR effector AVR3a is required for its stability in host cells to modulate plant immunity. Proc. Natl. Acad. Sci. U. S. A. *108*, 14682–14687.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. *11*, 367–372.

Zheng, W., Huang, L., Huang, J., Wang, X., Chen, X., Zhao, J., Guo, J., Zhuang, H., Qiu, C., Liu, J., et al. (2013). High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus. Nat. Commun. *4*:2673.